sensors

MDPI

Article

Deep Learning-Based Speech Enhancement of an Extrinsic Fabry–Perot Interferometric Fiber Acoustic Sensor System

Shiyi Chai [1,2], Can Guo [1,2], Chenggang Guan [2] and Li Fang [1,*]

[1] School of Science, Hubei University of Technology, Wuhan 430068, China
[2] Hubei Engineering Technology Research Center of Energy Photoelectric Device and System, Hubei University of Technology, Wuhan 430068, China
* Correspondence: fangli@hbut.edu.cn

Abstract: To achieve high-quality voice communication technology without noise interference in flammable, explosive and strong electromagnetic environments, the speech enhancement technology of a fiber-optic external Fabry–Perot interferometric (EFPI) acoustic sensor based on deep learning is studied in this paper. The combination of a complex-valued convolutional neural network and a long short-term memory (CV-CNN-LSTM) model is proposed for speech enhancement in the EFPI acoustic sensing system. Moreover, the 3 × 3 coupler algorithm is used to demodulate voice signals. Then, the short-time Fourier transform (STFT) spectrogram features of voice signals are divided into a training set and a test set. The training set is input into the established CV-CNN-LSTM model for model training, and the test set is input into the trained model for testing. The experimental findings reveal that the proposed CV-CNN-LSTM model demonstrates exceptional speech enhancement performance, boasting an average Perceptual Evaluation of Speech Quality (PESQ) score of 3.148. In comparison to the CV-CNN and CV-LSTM models, this innovative model achieves a remarkable PESQ score improvement of 9.7% and 11.4%, respectively. Furthermore, the average Short-Time Objective Intelligibility (STOI) score witnesses significant enhancements of 4.04 and 2.83 when contrasted with the CV-CNN and CV-LSTM models, respectively.

Keywords: optical fiber sensor; external Fabry–Perot interferometer; speech enhancement; CV-CNN; CV-LSTM

Citation: Chai, S.; Guo, C.; Guan, C.; Fang, L. Deep Learning-Based Speech Enhancement of an Extrinsic Fabry–Perot Interferometric Fiber Acoustic Sensor System. *Sensors* **2023**, *23*, 3574. https://doi.org/10.3390/s23073574

Academic Editor: Nélia J. Alberto

Received: 3 March 2023
Revised: 17 March 2023
Accepted: 23 March 2023
Published: 29 March 2023

1. Introduction

It is necessary to achieve high-quality voice communication in environments that have high temperatures, high pressure, strong radiation and strong electromagnetic effects. These environments create difficulties for conventional electroacoustic sensors and cause them not to work properly [1]. External Fabry–Perot interferometric (EFPI) acoustic sensors are widely used in special fields such as national defense and security [2], marine acoustic monitoring and positioning [3] and fuel pipeline leakage and positioning [4] because of their passive detection end, anti-electromagnetic interference, low loss, corrosion resistance and long-distance capabilities [5–7]. However, since the noise present in these environments can significantly degrade the perceptual quality and clarity of voice communication, voice enhancement is a much-needed task.

Speech enhancement is one of the most important and challenging tasks in speech applications, in which the goal is to suppress and reduce noise interference to extract useful speech signals in noisy backgrounds [8,9]. With the successful application of deep learning in the field of images [10–14], many scholars have begun to apply deep learning technology to speech enhancement. The existing speech enhancement methods can be divided into two categories: machine learning and deep learning. Regarding machine learning, early algorithms were generally implemented on shallow models and small datasets due to the limitations of computer hardware. Kim et al. developed a Gaussian mixture model (GMM)-based method for time–frequency (T-F) units according to the frequency band

characteristics of input signals to determine the probability of speech and noise. However, this method models each frequency band separately, ignoring the correlation between frequency bands [15]. Han et al. used the support vector machine (SVM) method to identify speech-dominated and noise-dominated T-F units. Compared with the GMM, the SVM shows a better generalization ability, but it still loses some target speech when the noise speech energy is too high [16]. Chung et al. designed a training and compensation algorithm of class-conditioned basis vectors in a nonnegative matrix factorization (NMF) model for single-channel speech enhancement. The NMF algorithm is trained separately on clean speech and noise to reduce residual noise components that have similar characteristics to clean speech. However, when encountering speech or noise that does not appear in training, the performance of the algorithm will drop [17].

Regarding deep learning, neural networks can enhance noisy speech in the time–frequency (TF) domain or directly in the time domain. In the time-domain method, the neural network directly learns the mapping relationship of the time-domain waveform level, and the processing flow is relatively simple. A one-dimensional convolutional neural network is usually used. Between feature extraction and waveform restoration, a neural network with a temporal modeling ability, such as a recurrent neural network (RNN) [18] and a temporal convolutional network (TCN) [19], is used to enhance the effect of speech enhancement. However, a shared limitation of RNNs and TCNs is their inability to effectively capture long-term dependencies in speech signals that extend across multiple time frames. On the other hand, TF domain methods, which are also popular, perform speech enhancement by learning the masking relationship from the spectrogram of noises to the spectrogram of clean speech. It is believed that speech and noise are more separated after passing through the short-time Fourier transform (STFT), which usually uses a convolutional encoder–decoder (CED) or U-Net framework [20]. After the STFT was performed, TF domain methods can take the complex-valued spectrogram as input and then decompose the magnitude and phase into real and imaginary parts in Cartesian coordinates. These methods solve the long-standing problem of the phase being difficult to estimate. Recently, Choi et al. proposed deep complex U-Net (DCUnet), which utilizes deep complex-valued convolutional layers for the near-perfect enhancement of speech [21]. Cao et al. proposed a generative adversarial network to model temporal and frequency correlations and achieved extremely high performance [22]. Park et al. proposed a multi-view attention network to improve the accuracy of feature extraction [23].

The EFPI acoustic sensor is a special microphone that converts sound waves into an optical signal. The speech signal needs to be received by the EFPI acoustic sensor and demodulated by the demodulation system to restore the sound signal. Unlike conventional microphones, some optical parameters (such as the light intensity, phase and polarization state) will also affect the acoustic characteristics of the EFPI acoustic sensor and cause frequency responses and noise [24,25]. Although deep learning has achieved extremely high performance in speech enhancement, most speech enhancement models are designed for conventional microphones and may not work perfectly on EFPI acoustic sensors.

In this paper, we propose a hybrid deep learning architecture that combines a complex-valued convolutional neural network and a long short-term memory (CV-CNN-LSTM) for speech enhancement of fiber-optic EFPI acoustic sensors. The CV-CNN and LSTM are set as the primary framework for the neural network, and the TF domain features of the signal are extracted through the STFT, completing the purpose of speech enhancement. Compared with a simple CV-CNN and LSTM, CV-CNN-LSTM is the best with regard to the speech enhancement performance test.

2. Basic Configuration and Data Acquire

2.1. Basic Configuration

Our EFPI acoustic sensor system with three-wavelength demodulation is schematically shown in Figure 1. An Er-doped amplified spontaneous emission (ASE) broadband source with an output power of 100 mw was used in our experiment. The output light from

the ASE source is incident to the EFPI sensor head through an optical circulator. The EFPI sensor head is formed by the cleaved end face of a single-mode-fiber (SMF) and silicon nitride diaphragm, which creates two reflective mirrors in the EFPI cavity. The cavity length of our EFPI sensor is approximately 100 μm. When an acoustic wave is applied to the silicon nitride diaphragm, it vibrates with the applied sound pressure, which modulates the cavity length of the EFPI and consequently induces the phase change of the interferential output light. The reflected light beams modulated with the phase signal are incident to a wavelength division multiplexer (WDM) through the optical circulator and divided into three beams according to their wavelengths. The wavelength interval of the three wavelengths is chosen according to the free spectrum range (FSR) of our EFPI sensor head. Each beam is then collected by a photodetector (PD) and converted into a voltage signal. The voltage signals are collected by data acquisition (DAQ) and processed by a computer. A loudspeaker excited by an audio analyzer is used as the acoustic source, which can generate sinusoidal acoustic waves at a specific frequency.

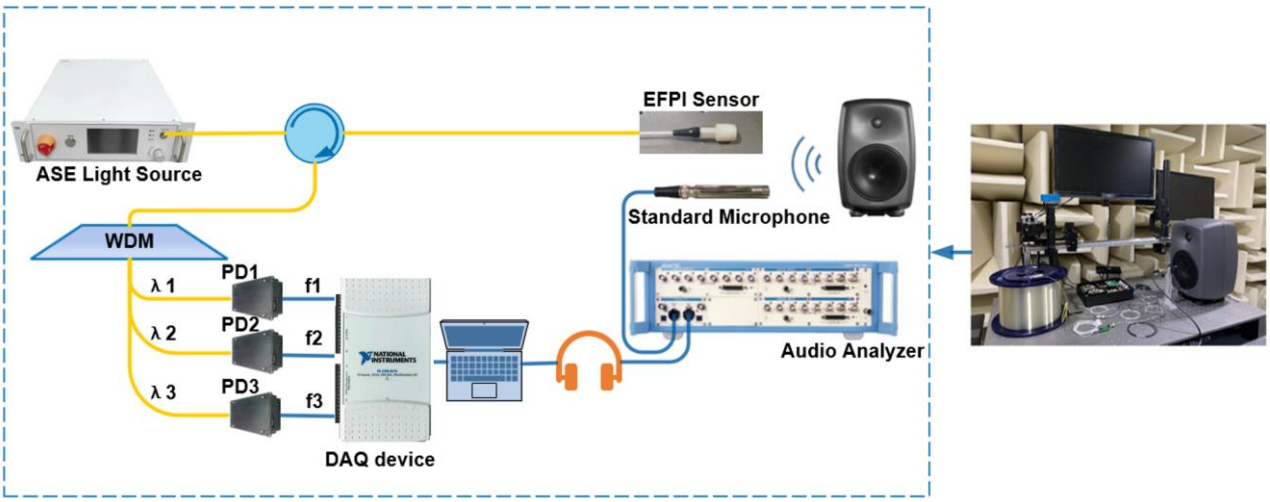

Figure 1. Schematic of the three-wavelength EFPI acoustic sensor system.

It is constructed entirely within a silent room to guarantee the system remains unaffected by external conditions during testing. This provides an environment of utmost silence to eliminate or minimize the impact of external noise and echoes on the recorded or tested audio.

2.2. Demodulation Principle

The intensity of the reflected interferential light at the three quadrature wavelengths can be expressed as

$$\begin{matrix} I_i = A + Bcos(\varphi_i) \\ \varphi_i = \frac{4n\pi}{\lambda_i} d_t \end{matrix}, (i = 1,\ 2,\ 3) \tag{1}$$

where λ_i is the output wavelength, A is the DC component of the interferometric fringe, B is the interferometric fringe visibility, n is the refractive index of the EFPI cavity, n = 1 and d_t is the cavity length modulated by the vibration signal. The relationship of the three wavelengths is calculated as follows to satisfy the quadrature relationship between the three output wavelengths.

$$\begin{matrix} 4\pi L/\lambda_1 + 2\pi/3 = 4\pi L/\lambda_2 \\ 4\pi L/\lambda_1 + 4\pi/3 = 4\pi L/\lambda_3 \end{matrix} \tag{2}$$

where L is the initial cavity length.

Considering that the three wavelengths cannot be completely equal in power and the phase difference cannot strictly meet $2\pi/3$, the three interferometric signals received by the DAQ can be described by

$$\begin{cases} f_1 = D_1 + E_1 \cos[\phi(t) + 2\pi/3 + \varphi_1] \\ f_2 = D_2 + E_2 \cos[\phi(t) + \varphi_2] \\ f_3 = D_3 + E_3 \cos[\phi(t) - 2\pi/3 - \varphi_3] \end{cases} \tag{3}$$

where D_1, D_2 and D_3 are the DC components of the interferometric fringe. E_1, E_2 and E_3 are the fringe visibility. φ_1, φ_2 and φ_3 are the phase deviations of the three outputs. $\phi(t)$ is the external disturbance signal, and t represents the time. A new output can be obtained by taking the average value of the two signals f_1 and f_2 output in the asymmetric state, and the new output p_1 can be expressed as

$$p_1 = (D_1 + D_2)/2 + E_4 \cos[\phi(t) + 2\pi/3 + \varphi_4] \tag{4}$$

where E_4 is the interference fringe visibility of the new signal p_1, φ_4 is the phase deviation of p_1 and the DC component and phase deviation are the most critical factors affecting the 3×3 coupler algorithm [26]. Similarly, by performing the same operation on f_2 and f_3, another new signal, P2 can be expressed as

$$p_2 = (D_2 + D_3)/2 + E_5 \cos[\phi(t) + 2\pi/3 + \varphi_5] \tag{5}$$

where E_5 is the DC component of p_2, φ_5 is the phase deviation of p_2. Moreover, p_1, p_2 and f_2 are used as new inputs to the 33 coupler algorithm. It can be seen from the three new signals that the difference between the DC components is very small after the operation. Compared with the original signals, the DC components of the three new signals are closer in value. In addition, the phase deviation is relatively small, and it is compressed to be smaller after the operation. Thus, the errors caused by the obtained φ_4 and φ_5 can be ignored.

2.3. Feature Extraction

The short-time Fourier transform (STFT) is used to extract the time–frequency domain features of speech signals. The STFT has excellent time–frequency resolution, which means that it can accurately localize changes in the frequency content of a signal over time. This makes it a powerful tool for analyzing time-varying signals. For a discrete signal $x(n)$ of length N, the discrete STFT at the frequency f and the short interval at the tth moment can be expressed as follows

$$X(f,t) = \sum_{n=0}^{N-1} x(n)w(n - tB)e^{-j2\pi fn} \tag{6}$$

where $w(n)$ is the window function and B is the hop length. The spectrogram representation of speech data is influenced by both the window size and the hop length. Specifically, the window size primarily affects the frequency resolution, while the time resolution is primarily influenced by the hop length. Figure 2 shows the spectrograms of a clean speech signal and a noisy speech signal.

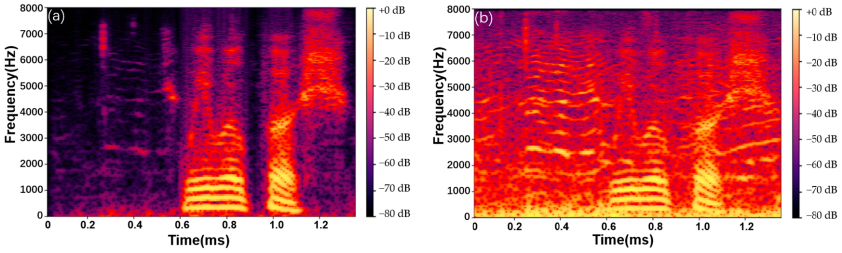

Figure 2. The spectrogram of (**a**) a clean speech signal and (**b**) a noisy speech signal.

3. Neural Networks

3.1. Complex-Valued Convolution

Complex-valued convolution is an operation in mathematics and signal processing in which two complex-valued functions are convolved with each other. A complex-valued spectrogram obtained by the STFT of a speech signal can be decomposed into real and imaginary parts in Cartesian coordinates. The CV-CNN convolves the real part and imaginary part, respectively, with two complex-valued functions. The complex-valued convolution filter, also known as kernels, is defined as $\boldsymbol{W} = \boldsymbol{A} + \boldsymbol{iB}$, where both $\boldsymbol{A}$ and $\boldsymbol{B}$ are real-valued matrices. The input complex matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \boldsymbol{X_r} + \boldsymbol{iX_i}$, and the complex-valued convolution operation on $\boldsymbol{W}$ with $\boldsymbol{X}$ is done by $\boldsymbol{W} * \boldsymbol{X} = (\boldsymbol{A} * \boldsymbol{X_r} - \boldsymbol{B} * \boldsymbol{X_i}) + \boldsymbol{i}(\boldsymbol{B} * \boldsymbol{X_r} + \boldsymbol{A} * \boldsymbol{X_i})$. The operation process of complex-valued convolution is shown in Figure 2.

The complex-valued convolutional layers constitute the main structural elements of a CV-CNN and extract features from the complex-valued input data. A complex-valued convolution can be implemented as a set of four convolutions with real-valued functions. The two complex-valued functions are separated into their real and imaginary components, and each component is convolved separately with the corresponding component of the other function. Thus four real-valued convolutions can be combined to form the final complex-valued convolution. These convolution kernels are two-dimensional (2D) since the one-dimensional (1D) speech time waveforms have been transformed into a complex-valued spectrogram with the application of the STFT. Since the spectrogram is complex-valued, the dot product is computed separately for the real and imaginary parts of the filter and the spectrogram, resulting in a complex-valued output at each position.

Multiple hidden layers are included in a typical complex-valued convolutional structure, which can extract features of high-dimensional data and learn nonlinear relationships adaptively. We add a batch normalization (BN) layer and a Leaky ReLu (LR) layer after each convolutional layer. The BN layer can improve the training stability and convergence speed of the network by normalizing the activations of the previous layer. When used after each convolutional layer, the BN layers can help to stabilize the training process by maintaining the variance and mean of the activations close to zero and one, respectively. The LR activation functions are a variant of the ReLu that allows a small gradient when the input is negative. The LR layer can help to introduce nonlinearity and improve the ability of the network to learn complex representations of the input data. A complete complex-valued convolutional network structure is shown in Figure 3.

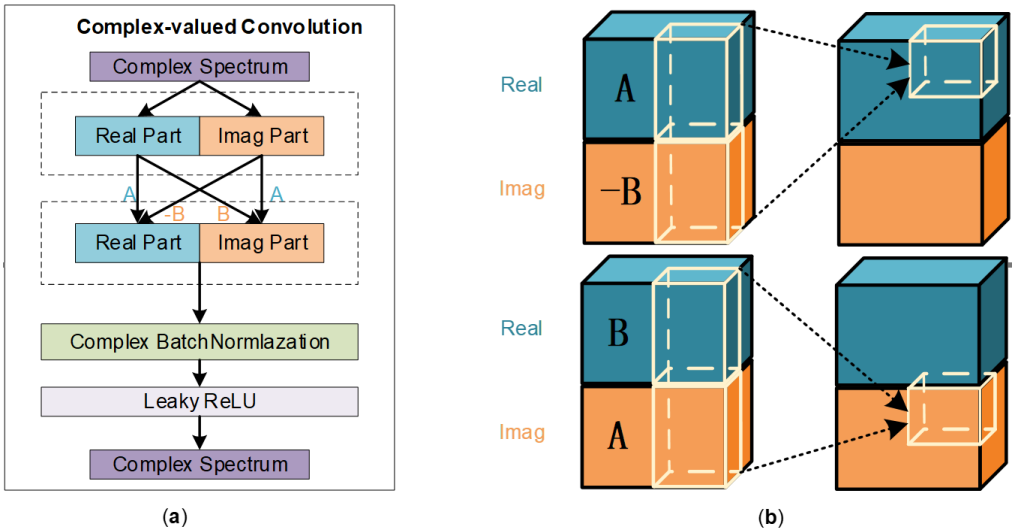

Figure 3. (**a**) The specific structure of complex-valued convolution. (**b**) Calculation of real and imaginary parts in complex-valued convolution operations.

3.2. Long Short-Term Memory

When deep learning is applied to sequential data such as music, video and speech, it is important to model the long-term dependencies in the time series. RNNs are designed to process sequential data using the mechanism of recurrent feedback. However, RNNs are notoriously difficult to train due to the vanishing and exploding gradient issues. To overcome these problems, the LSTM was proposed as a special type of RNN. There are three important components inside the LSTM, including the input gate, forgetting gates and output gates. These gates allow the network to selectively retain or discard information from previous time steps, ensuring that relevant features of the sequential data are accurately captured. By selectively controlling the flow of information through the network, LSTMs can effectively model long-term dependencies and have become a widely used approach in various applications of sequential data analysis.

CV-CNN excels at learning time–frequency domain features in speech signals, but its ability to capture time dependence and long-term context information is limited. In contrast, LSTMs are adept at capturing long-term dependencies in speech signals, which may span across multiple time frames. This makes them well-suited for modeling the temporal structure of speech and reducing the impact of noise interference. By integrating LSTM with CV-CNN, we can build a model that effectively handles speech signals of different lengths without requiring extensive preprocessing. In order to be able to handle complex-valued features extracted by the CV-CNN, we use the CV-LSTM in the proposed model. Similar to CV-CNN, considering the real and imaginary components of the complex input X_r and X_i, the output of the CV-LSTM, F_{out}, can be defined as

$$\begin{cases} F_{rr} = \mathrm{LSTM_r}(X_r); F_{ir} = \mathrm{LSTM_r}(X_i) \\ F_{ri} = \mathrm{LSTM}_i(X_r); F_{ii} = \mathrm{LSTM}_i(X_i) \\ F_{out} = (F_{rr} - F_{ii}) + j(F_{ri} + F_{ir}) \end{cases} \tag{7}$$

where $\mathrm{LSTM_r}$ and LSTM_i represent two traditional LSTMs of the real part and the imaginary part.

3.3. Target and Loss Function

The target of model training is a complex ratio mask (CRM). CRM is a technique used in speech enhancement. It is a complex-valued function that estimates the ratio between the desired speech signal and the interfering noise signal. The CRM is computed by taking the complex ratio of the time–frequency representations of the desired speech and the noisy signal. The resulting CRM is then applied to the noisy signal to suppress the noise and enhance the desired speech signal. The CRM is able to capture the phase information of the signal, which is useful when the noise around the EFPI acoustic sensor is unstable, and its phase changes over time; CRM can be defined as

$$CRM = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + j\frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2} \tag{8}$$

wre Y_r and Y_i denote the real and imaginary parts of the noisy complex-valued spectrogram obtained by the STFT, respectively. The real and imaginary parts of the clean speech complex-valued spectrogram obtained by the STFT are represented by S_r and S_i.

The loss function measures the discrepancy between the predicted output of the model and the true output, and the goal of the model is to minimize this discrepancy. In the present study, the scale-invariant signal-to-noise ratio (SI-SNR) is utilized. The SI-SNR is a metric commonly used to evaluate the performance of speech separation or source separation algorithms. It measures the ratio of the energy of the target speech signal to the

energy of the interference or noise signal while being insensitive to the amplitude scaling of the separated signals. The SI-SNR can be defined as

$$\mathrm{SI}-\mathrm{SNR}=10*\log_{10}(\frac{||\mathrm{s}||^2}{||\mathrm{s}-\mathrm{y}||^2}) \tag{9}$$

where s is the reference signal and y is the estimated signal; the $||.||$ operator denotes the L2 norm.

3.4. The CV-CNN-LSTM Model

In the present study, the CV-CNN-LSTM model mainly adopts the encoder–decoder structure. The encoder receives input and compresses it into a reduced representation, which is subsequently forwarded to a decoder. The decoder then generates an output based on this compressed representation. The encoder–decoder framework is commonly used in applications such as natural language processing, speech and audio processing and image and video processing.

In the CV-CNN-LSTM model, the encoder is the complete complex-valued convolution network structure mentioned in Section 3.1. This encoder is composed of a complex-valued convolution layer, batch normalization layer and LeakyReLu layer. The decoder is similar in structure to the encoder, except that all 2D convolution functions are replaced by 2D transposed functions. The structure of the CV-CNN-LSTM model is shown in Figure 4. This model consists of six encoder blocks, six decoder blocks and one CV-LSTM) layer. Moreover, the FC is a fully connected layer. The fully connected layers in the model are used to learn nonlinear combinations of features at a higher level of abstraction.

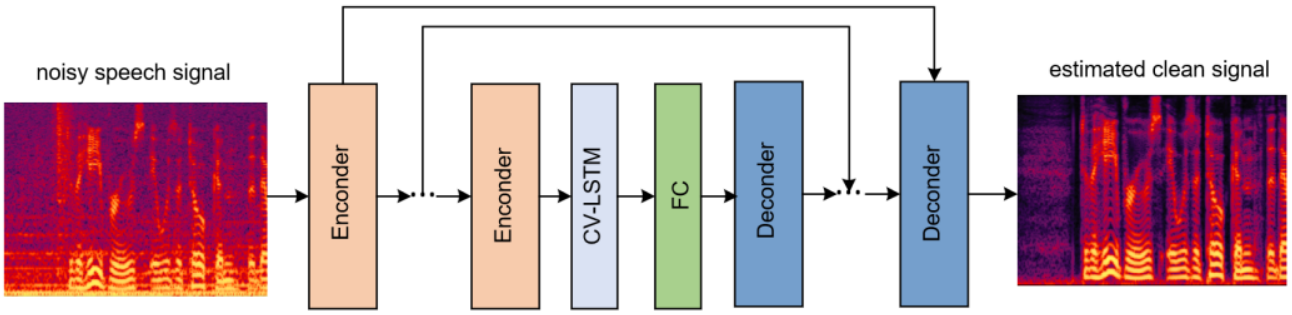

Figure 4. The structure of CV-CNN-LSTM model for speech enhancement.

To improve the performance and training efficiency of the encoder–decoder architecture, we use skip connections between the encoder and decoder. Skip connections enable the network to reuse learned features from earlier layers in later layers, which can help to preserve important information and prevent that information from being lost during training. Additionally, with skip connections, the network can converge faster because it is easier to learn identity maps than to learn complex maps from scratch.

4. Results and Discussion

To ensure the normal demodulation of the optimized 3 × 3 algorithm, we first conducted a performance test in an anechoic chamber in which the three wavelengths are selected according to Equation (2): 1546.92 nm (λ_1), 1550.92 nm (λ_2) and 1554.94 nm (λ_3). The NI-USB6210 is used to collect and process the signals of the three PDs. The demodulated signals are connected to the audio analyzer through the earphones for testing. The demodulation effects of the different frequency signals are shown in Figure 5. To obtain more detailed indicators, we used the audio analyzer test to evaluate the signal-to-noise ratio. The results, which reach a signal-to-noise ratio of 62 DB, are shown in Figure 6.

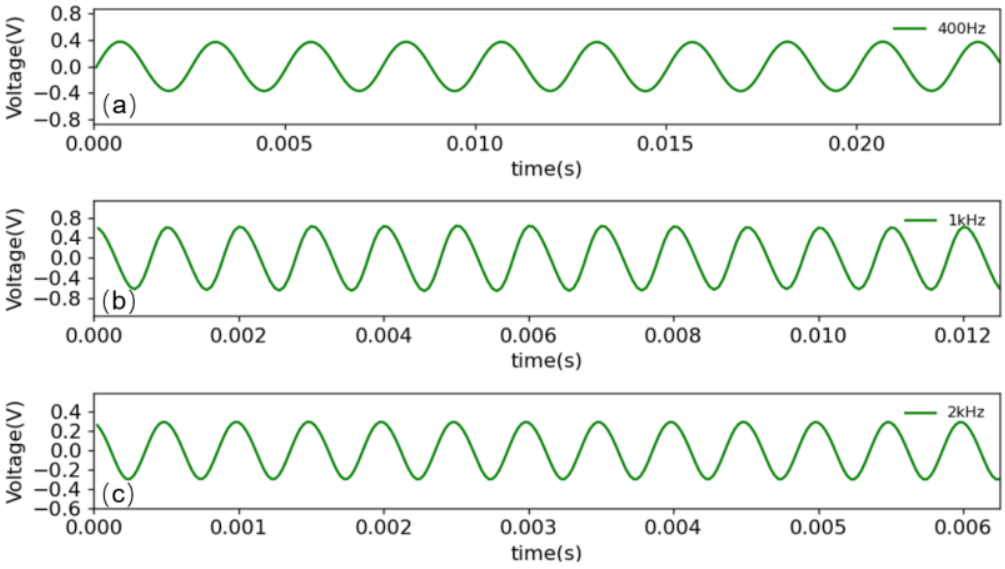

Figure 5. Demodulation results at different frequencies: (**a**) the demodulated signal of 400 Hz, (**b**) the demodulated signal of 1 KHz, (**c**) the demodulated signal of 2 KHz.

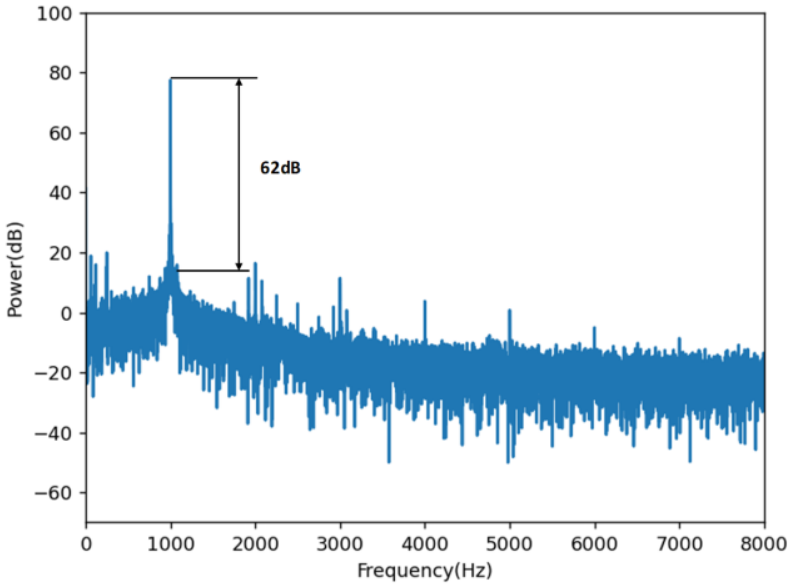

Figure 6. The result of the signal-to-noise ratio.

In this study, our dataset is constructed from the 28 speakers dataset (28spk) [27], which contains speech clips from 28 people. We randomly selected 6000 utterances from the 28spk corpus and divided them into three parts. There were 4800 utterances in the training set, 660 utterances in the validation set and 540 utterances in the test set. All utterances were played in the anechoic chamber through a high-quality speaker, and EFPI acoustic sensors were used to capture the speech signals. In the end, the audio was clipped to ten seconds, and we obtained approximately 16 h of paired clean and noisy utterances. The evaluation set is generated by randomly selecting utterances from the speech set and the noise set and mixing them at three signal-to-noise ratios (SNRs) (5 DB, 10 DB and 15 DB). According to the frequency range of the voice call, all the speech signals were sampled at 16 kHz. The design and implementation of the deep learning model for the speech enhancement of the EFPI acoustic sensor were held in a Python 3.8.3 environment using the deep learning tool Torch. Pytorch 1.7 was used as the backend of the Torch library.

All experiments were performed on a desktop computer featuring an Intel Core i7-10700 2.90 GHz CPU, 32 GB RAM memory and a 10 GB NVIDIA GeForce RTX 3080 GPU.

Regarding the CV-CNN-LSTM model, the complex-valued spectrogram was extracted by using a Hamming window. The window length and hop size are 25 ms and 6.25 ms, respectively, and the FFT length is 512. The number of channels for the CV-CNN-LSTM is {32, 64, 128, 256, 256, 256}. The kernel size and stride are set to (5, 2) and (2, 1), respectively. The CV-LSTM layer uses a two-layer structure, the parameters of the two layers are the same, and each layer contains 128 hidden units. We chose Adam as the optimizer, set the initial learning rate to 0.001 and used ExponentialLR to control the change in the learning rate. We compare several models, including CNN, LSTM, CV-CNN and CV-LSTM, on the same dataset. The CNN model is structured with six 2D convolutional layers, each accompanied by a batch normalization layer and succeeded by a max-pooling layer. Channel counts for the CNN model are specified as {16, 32, 64, 64, 128, 128}, while kernel sizes and strides are consistently set to (3, 3) and (1, 1). The LSTM model contains two LSTM layers; each layer has 256 units, and the output layer is a fully-connected layer. Comprising six CV-CNN layers, the CV-CNN model boasts channel numbers of {16, 32, 64, 64, 128, 128} and identical kernel size and stride settings as its CNN counterpart. Lastly, the CV-LSTM model is composed of two CV-LSTM layers, and each LSTM layer has 256 units and utilizes two separate fully-connected layers to deliver the real and imaginary components of the results, respectively.

The evaluation of speech enhancement model performance is conducted using two widely-accepted and complementary metrics: Perceptual Evaluation of Speech Quality (PESQ) [28] and Short-Time Objective Intelligibility (STOI) [29]. PESQ scores generally range from -0.5 to 4.5, with higher values indicating superior speech quality. The STOI metric provides a normalized score between 0 and 1, where higher values correspond to greater speech intelligibility. Tables 1 and 2 show the comparison of PESQ and STOI scores between the proposed model and other models. It can be seen that the proposed model has the best test results on data with different SNRs.

Table 1. The PESQ scores of different models at different SNRs.

Model	5 dB	10 dB	15 dB	Ave.
CNN	2.347	2.544	2.627	2.506
CV-CNN	2.724	2.835	3.045	2.868
LSTM	2.701	2.846	2.928	2.825
CV-LSTM	2.746	2.897	3.107	2.917
CV-CNN-LSTM	2.948	3.135	3.361	3.148

Table 2. The STOI (in %) scores of different models at different SNRs.

Model	5 dB	10 dB	15 dB	Ave.
CNN	77.97	83.15	85.08	82.07
CV-CNN	86.73	87.65	91.97	88.78
LSTM	86.59	87.71	89.35	87.88
CV-LSTM	86.94	88.61	92.42	89.99
CV-CNN-LSTM	89.51	93.14	95.83	92.82

Furthermore, additional experiments were conducted to optimize the proposed model. In particular, the performance of the proposed model was evaluated while changing some of its parameters, namely, the window length of the STFT and the number of complex-valued convolutional layers. The window length is a key parameter in the STFT for the feature extraction of speech signals. Using a window length that is too long can result in poor time resolution. This results in a loss of important spectral information and increased spectral leakage. Conversely, if the window length is too short, it will result in a poor frequency resolution. This can lead to poor separation of the speech and noise components in the frequency domain, resulting in a low quality of enhanced speech. The experimental

results of the window length comparison are shown in Tables 3 and 4. It can be seen that the optimal window length is approximately 25 ms. The enhancement effect of the CV-CNN-LSTM model that uses different window lengths is shown in Figure 7. The noise signal used in Figure 6 is a speech recording at an SNR of 10 dB.

Table 3. The PESQ scores for models with different window lengths.

Win_Length (ms)	5 dB	10 dB	15 dB	Ave.
64	2.619	2.826	3.049	2.831
40	2.801	2.904	3.212	2.972
25	2.948	3.135	3.361	3.148
15	2.745	2.843	3.108	2.899

Table 4. The STOI (in %) scores for models with different window lengths.

Win_Length (ms)	5 dB	10 dB	15 dB	Ave.
64	84.91	87.52	91.98	88.14
40	87.29	88.76	94.51	90.19
25	89.51	93.14	95.83	92.82
15	86.93	87.61	92.38	88.97

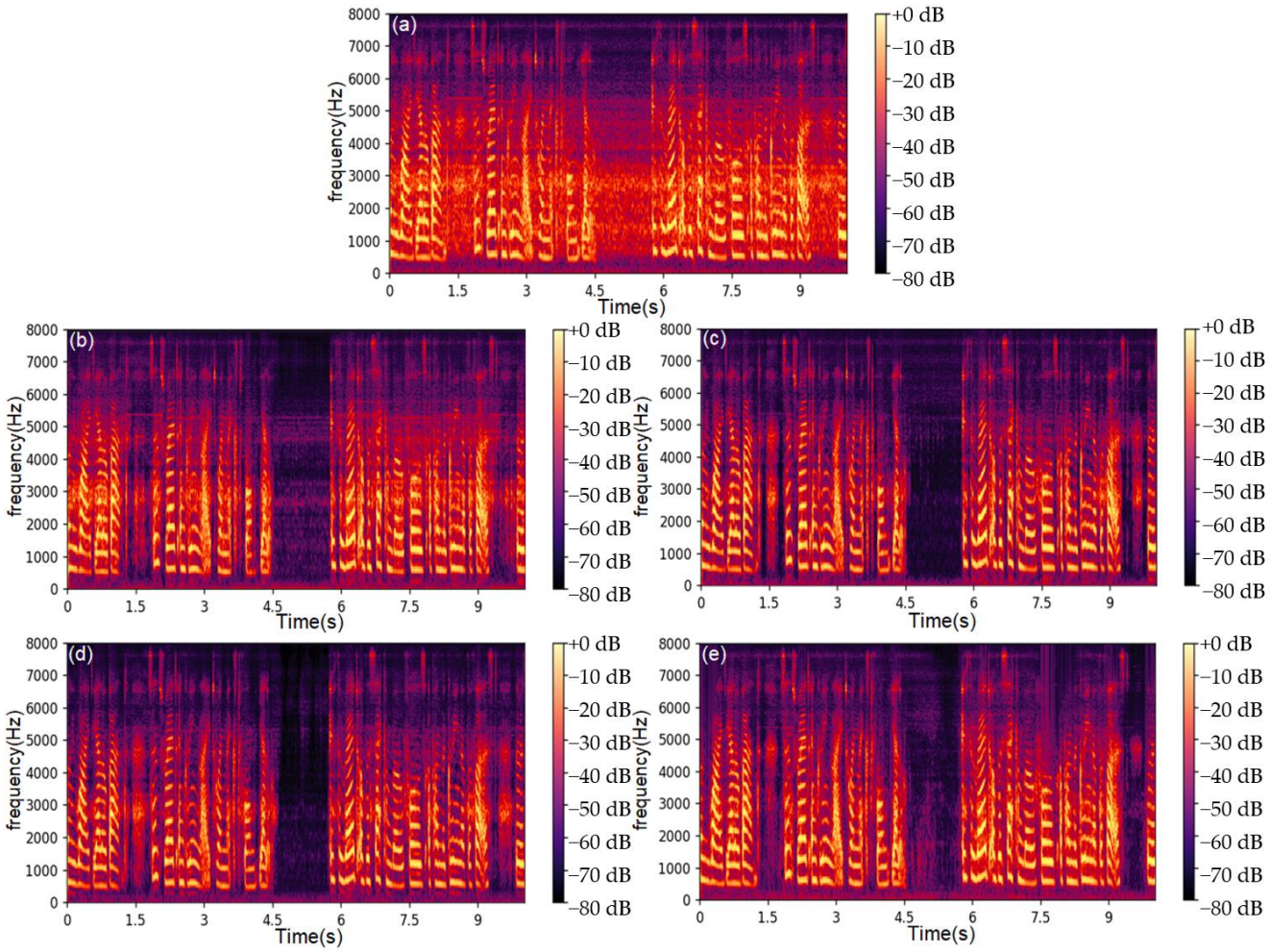

Figure 7. Test results of models with different window lengths: (**a**) STFT spectrogram of original speech signal, (**b**) model with a window length of 15 ms, (**c**) model with a window length of 25 ms, (**d**) model with a window length of 40 ms, (**e**) model with a window length of 64 ms.

The number of complex-valued convolutional layers also plays a significant role in the model's performance. If the number is too low, the model may fail to capture crucial input data features effectively. On the other hand, excessive layers may lead to overfitting, resulting in good training set scores but poor performance on the test set. Tables 5 and 6 show the performance of the proposed model for different numbers of complex-valued convolutional layers. The model was trained and tested using four to eight complex-valued convolutional layers in a structure similar to the one displayed in Figure 4. As can be observed, both PESQ and STOI scores are highest when using six complex-valued convolutional layers in the CV-CNN-LSTM model.

Table 5. The PESQ scores for models with different complex-valued convolutional layers.

No. of Complex-Valued Convolutional Layers	5 dB	10 dB	15 dB	Ave.
4	2.768	2.921	3.126	2.938
5	2.813	3.019	3.205	3.012
6	2.948	3.135	3.361	3.148
7	2.804	2.977	3.156	2.979
8	2.932	3.044	3.275	3.083

Table 6. The STOI (in %) scores for models with different complex-valued convolutional layers.

No. of Complex-Valued Convolutional Layers	5 dB	10 dB	15 dB	Ave.
4	87.34	89.27	93.05	92.87
5	87.28	91.63	94.22	91.04
6	89.51	93.14	95.83	92.82
7	87.16	91.37	93.49	90.67
8	89.42	91.28	94.55	91.75

5. Conclusions

In this paper, speech enhancement techniques based on fiber-optic EFPI acoustic sensors are studied. First, the speaker's speech signal is demodulated by the fiber-optic EFPI acoustic sensor demodulated based on the 3×3 coupling algorithm, and then the speech signal is edited, and the edited speech signal is subjected to STFT to extract spectral features. The overall structure of the CV-CNN-LSTM model is implemented by combining CV-CNN and CV-LSTM. Among them, CV-CNN is suitable for processing complex-valued spectrogram data, while CV-LSTM is good at capturing the characteristics of sequential data related to time series. Experimental results show that the CV-CNN-LSTM model can achieve better performance than other models in terms of PESQ score and STOI score.

The speech enhancement technology in this paper is expected to be applied to the fields where traditional methods cannot be applied, such as high magnetic field environments, flammable and explosive environments and high electric field environments. Of course, the fiber-optic EFPI acoustic sensor process in this paper is more complex than electrical-based acoustic sensors. With the reduction of the cost of optoelectronic devices, the technology is expected to be used in fields such as deserts and polar regions.

Author Contributions: Conceptualization, S.C. and C.G. (Can Guo); methodology, S.C. and C.G. (Can Guo); software, S.C. and C.G. (Can Guo); validation, S.C. and C.G. (Chenggang Guan); formal analysis, C.G. (Can Guo); investigation, S.C. and C.G. (Can Guo); resources, L.F. and C.G. (Chenggang Guan); data curation, C.G. (Can Guo); writing—original draft preparation, S.C. and C.G. (Can Guo); writing—review and editing, all authors; visualization, S.C. and C.G. (Can Guo); project administration, L.F. and C.G. (Chenggang Guan); All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project of outstanding young and middle-aged science and technology innovation teams of colleges and universities in Hubei province under Grant (No. T201907) and the International Science and Technology Cooperation Key Research and Development Program of Science and Technology Agency in Hubei Province (No. 2021EHB018).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Xiang, Z.; Dai, W.; Rao, W. A Gold Diaphragm-Based Fabry-Perot Interferometer with a Fiber-Optic Collimator for Acoustic Sensing. *IEEE Sens. J.* **2021**, *21*, 17882–17888. [CrossRef]
2. Wang, Y.; Yuan, H.; Liu, X. A comprehensive study of optical fiber acoustic sensing. *IEEE Access* **2019**, *7*, 85821–85837. [CrossRef]
3. Zhang, W.; Chen, F.; Ma, W. Ultrasonic imaging of seismic physical models using a fringe visibility enhanced fiber-optic Fabry-Perot interferometric sensor. *Opt. Express* **2018**, *26*, 11025–11033. [CrossRef] [PubMed]
4. Liu, Q.; Jing, Z.; Liu, Y. Multiplexing fiber-optic Fabry–Perot acoustic sensors using self-calibrating wavelength shifting interferometry. *Opt. Express* **2019**, *27*, 38191–38203. [CrossRef] [PubMed]
5. Jo, W.; Akkaya, O.C.; Solgaard, O.; Digonnet, M.J.F. Miniature fiber acoustic sensors using a photonic-crystal membrane. *Opt. Fiber Technol.* **2013**, *19*, 785–792. [CrossRef]
6. Zhang, W.; Lu, P.; Qu, Z. Passive Homodyne Phase Demodulation Technique Based on LF-TIT-DCM Algorithm for Interferometric Sensors. *Sensors* **2021**, *21*, 8257. [CrossRef] [PubMed]
7. Fu, X.; Lu, P.; Zhang, J. Micromachined extrinsic Fabry-Pérot cavity for low-frequency acoustic wave sensing. *Opt. Express* **2019**, *27*, 24300–24310. [CrossRef] [PubMed]
8. Chaudhari, A.; Dhonde, S.B. A review on speech enhancement techniques. In Proceedings of the 2015 International Conference on Pervasive Computing (ICPC), Pune, India, 8–10 January 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–3.
9. Michelsanti, D.; Tan, Z.H.; Zhang, S.X. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1368–1396. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
11. Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4340–4354. [CrossRef]
12. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph convolutional networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5966–5978. [CrossRef]
13. Ge, Z.; Liu, S.; Wang, F. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
14. Wu, X.; Hong, D.; Chanussot, J. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Trans. Image Process.* **2022**, *32*, 364–376. [CrossRef]
15. Kim, G.; Lu, Y.; Hu, Y. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *J. Acoust. Soc. Am.* **2009**, *126*, 1486–1494. [CrossRef] [PubMed]
16. Han, K.; Wang, D.L. Towards generalizing classification based speech separation. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *21*, 168–177. [CrossRef]
17. Chung, H.; Badeau, R.; Plourde, E. Training and compensation of class-conditioned NMF bases for speech enhancement. *Neurocomputing* **2018**, *284*, 107–118. [CrossRef]
18. Huang, P.S.; Kim, M.; Hasegawa-Johnson, M. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 2136–2147. [CrossRef]
19. Kishore, V.; Tiwari, N.; Paramasivam, P. Improved Speech Enhancement Using TCN with Multiple Encoder-Decoder Layers. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 4531–4535.
20. Tan, K.; Wang, D.L. A convolutional recurrent neural network for real-time speech enhancement. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3229–3233.
21. Choi, H.S.; Kim, J.H.; Huh, J. Phase-aware speech enhancement with deep complex u-net. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
22. Cao, R.; Abdulatif, S.; Yang, B. CMGAN: Conformer-based metric GAN for speech enhancement. *arXiv* **2022**, *preprint*. arXiv:2203.15149.
23. Park, H.J.; Kang, B.H.; Shin, W. Manner: Multi-view attention network for noise erasure. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7842–7846.
24. Lu, X.; Wu, Y.; Gong, Y. A miniature fiber-optic microphone based on an annular corrugated MEMS diaphragm. *J. Light. Technol.* **2018**, *36*, 5224–5229. [CrossRef]

25. Ge, Y.X.; Wang, M.; Yan, H.T. Mesa diaphragm-based Fabry-Perot optical MEMS pressure sensor. In Proceedings of the 2008 1st Asia-Pacific Optical Fiber Sensors Conference, Chengdu, China, 7–9 November 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–4.
26. Chiu, B.; Hastings, M.C. Digital demodulation for passive homodyne optical fiber interferometry based on a 3 by 3 coupler. In Proceedings of the Fiber Optic and Laser Sensors XII, San Diego, CA, USA, 24–29 July 1994; SPIE: Bellingham, WA, USA, 1994; Volume 2292, pp. 371–382.
27. 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; IEEE: Piscataway, NJ, USA, 2013; pp. 1–4.
28. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; Proceedings (Cat. No. 01CH37221). IEEE: Piscataway, NJ, USA, 2001; Volume 2, pp. 749–752.
29. 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]