applied sciences

MDPI

Article

Speech Enhancement Based on Two-Stage Processing with Deep Neural Network for Laser Doppler Vibrometer

Chengkai Cai [1,*], Kenta Iwai [2] and Takanobu Nishiura [2,*]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Kyoto 603-8577, Japan
[2] College of Information Science and Engineering, Ritsumeikan University, Kyoto 603-8577, Japan;
* Correspondence: gr0370hv@ed.ritsumei.ac.jp (C.C.); nishiura@is.ritsumei.ac.jp (T.N.)

Abstract: The development of distant-talk measurement systems has been attracting attention since they can be applied to many situations such as security and disaster relief. One such system that uses a device called a laser Doppler vibrometer (LDV) to acquire sound by measuring an object's vibration caused by the sound source has been proposed. Different from traditional microphones, an LDV can pick up the target sound from a distance even in a noisy environment. However, the acquired sounds are greatly distorted due to the object's shape and frequency response. Due to the particularity of the degradation of observed speech, conventional methods cannot be effectively applied to LDVs. We propose two speech enhancement methods that are based on two-stage processing with deep neural networks for LDVs. With the first proposed method, the amplitude spectrum of the observed speech is first restored. The phase difference between the observed and clean speech is then estimated using the restored amplitude spectrum. With the other proposed method, the low-frequency components of the observed speech are first restored. The high-frequency components are then estimated by the restored low-frequency components. The evaluation results indicate that they improved the observed speech in sound quality, deterioration degree, and intelligibility.

Keywords: distant-talking speech measurement; speech enhancement; deep neural network; laser Doppler vibrometer

Citation: Cai, C.; Iwai, K.; Nishiura, T. Speech Enhancement Based On Two-Stage Processing with Deep Neural Network for Laser Doppler Vibrometer. *Appl. Sci.* **2023**, *13*, 1958. https://doi.org/10.3390/app13031958

Academic Editor: Yat Sze Choy

Received: 28 November 2022
Revised: 16 December 2022
Accepted: 26 December 2022
Published: 2 February 2023

1. Introduction

Distant-talk measurement is useful in crime prevention, security, and disaster relief. A traditional microphone obtains an acoustic signal by converting the sound wave into electric signals through an internal diaphragm. However, since the energy of a sound wave attenuates as the distance increases when propagating in the air, it is difficult to acquire distant sounds using traditional microphones. To solve this problem, parabolic microphones and shotgun microphones have been developed [1]. These microphones can pick up remote voices by changing the shape and position of the diaphragm. When using these microphones to acquire remote sound, however, the noise around the microphone will also be picked up. Therefore, it is difficult to pick up only the distant target sound in a noisy environment. Based on the fact that a laser can reach over a long distance and its energy attenuates slowly, systems using a laser Doppler vibrometer (LDV) [2] to acquire sound have been proposed [3–5]. By emitting a laser beam to a vibrating object and receiving the reflected light, an LDV can compare the phase difference between the reflected light and reference light in real time to calculate the current vibration velocity of the object. Thus, the vibration of the sound source can be restored from the calculated vibration velocity. This technology is used in many fields such as industrial and medical measurements [6–8].

When acquiring sound using an LDV from an irradiated object, however, the observed speech is distorted by the vibration characteristics of the object. In other words, non-stationary noise due to the technical design of an LDV is mixed in the observed speech and

phase delay occurs due to the reflective and vibration characteristics of the irradiated object. Moreover, there is a lack of observed speech components in the small amplitude response frequency bands. Therefore, to obtain high-quality speech using an LDV, it is necessary to apply processing such as non-stationary noise suppression and high-frequency component restoration to the observed speech.

Speech enhancement methods for LDVs have been proposed [9–12]. Li et al. [9] developed a method of first using a band-pass filter to filter out the noise in the non-voice frequency band of the acquired speech and then using a Wiener filter to reduce the noise of the speech signal. Peng et al. [10] proposed acquiring speech with two LDVs then applying the coherent-to-diffuse ratio (CDR) and multi-channel linear prediction (MCLP) between the acquired two-channel signals for noise reduction. However, these methods only execute noise-reduction processing for the acquired speech, and any components missing due to the frequency response are not restored. Xie et al. [11] proposed a method for encoding the mapping between observed speech by using an LDV and clean speech as an auxiliary feature to improve the speech recognition accuracy in a noisy environment, but this method still does not solve the problem of degradation caused by the vibration characteristics of the irradiated object. In recent years, deep neural networks (DNNs) have shown excellent results in speech signal processing [13–17]. With these conventional speech enhancement methods, the mapping between the input power spectrum and target power spectrum can be learned through the network, and the phase of the observed speech is generally used when restoring the speech [13,14] or directly inputting the waveform into the network to learn the mapping of the observed speech and clean speech [15]. The research on phase is also gradually attracting attention. When the phase information is unknown, the initial phase spectrum can be learned from the power spectrum and then adjusted using the Griffin–Lim algorithm (GLA) [16,17]. Since the observed speech has various types of deterioration, conventional speech enhancement methods cannot be effectively applied directly to the observed speech from LDVs.

Considering the deterioration of observed speech, we propose two speech-enhancement methods that are based on the frequency and time domains, respectively. Both methods use multiple DNNs to handle different types of deterioration. With our frequency-domain-processing method (hereafter, short-time Fourier transfer (STFT)-based method), the noise power in the power spectrum of the observed speech is first removed. The phase difference between the observed speech and clean speech is then calculated using the noise-suppressed power spectrum to obtain the phase spectrum of the enhanced speech. With our time-domain-processing method (hereafter, waveform-based method), the low-frequency waveform is first restored, then the high-frequency waveform is estimated using the restored low-frequency waveform. We conducted experiments to evaluate the proposed methods, and the results indicate that both methods can effectively handle most types of deterioration in the observed speech from LDVs, an improvement over conventional speech enhancement methods.

This paper is organized as follows. Section 2 introduces the characteristics of observed speech from LDVs and the two proposed methods. Section 3 presents the evaluation experiments we conducted to evaluate the effectiveness of the proposed methods and the results. Section 4 concludes this paper and introduces future work.

2. Proposed Speech Enhancement Methods for LDVs

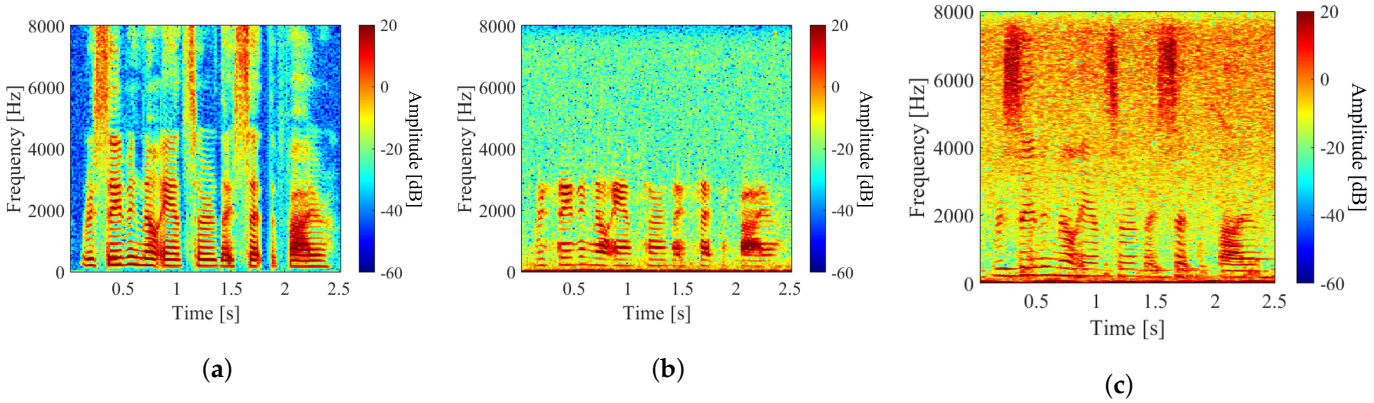
We first discuss the problems with LDVs when acquiring sounds then introduce the two proposed speech-enhancement methods, i.e., STFT- and waveform-based methods. The STFT-based method involves STFT-based two-stage processing of amplitude and phase reconstructions, and the waveform-based method involves waveform-based two-stage noise suppression processing and high-frequency component reconstruction. Each method is described in the following sections. All variables are listed in Table 1.

Table 1. List of notations.

Variable	Signification	Variable	Signification
$\boldsymbol{x}$	Clean speech	ϕ_x	Phase spectrum of clean speech
$\boldsymbol{y}$	Observed speech	ϕ_y	Phase spectrum of observed speech
x^{NB}	Narrow-band clean speech	ϕ^{PD}	Phase difference between ϕ_x and ϕ_y
y^{NB}	Narrow-band observed speech	K	Fourier transform length
x^{WB}	Wide-band clean speech	k	Frequency index
y^{WB}	Wide-band observed speech	m	Frame index
$\lvert \boldsymbol{X}^{\mathrm{LPS}} \rvert$	Log-power spectrum of clean speech	n	Sampling index
$\lvert \boldsymbol{Y}^{\mathrm{LPS}} \rvert$	Log-power spectrum of observed speech	$\hat{\text{value}}$	Estimated value

2.1. Problems with Sound Measurement Using LDVs

The sound wave can be restored by measuring the vibration generated by sound excitation. An LDV is a device that measures the moving velocity of an object by calculating the phase difference between the reference light and reflected light that irradiates the object surface [3]. Since a laser beam has high straightness, it is possible to measure vibration occurring far away and acquire only the sound near the irradiated object not affected by the noise surrounding the LDV. Therefore, the speech acquired from an LDV includes the frequency characteristics of the vibrating object. Therefore, the sound quality of observed speech greatly depends on the shape and vibration characteristics of the object. For example, objects with rough surfaces will cause tiny changes in the intensity and direction of reflected light. This also causes noise to mix into the observed speech. The vibration characteristics can also result in a lack of speech components in a small amplitude and phase delay in the frequency spectrum, as shown in Figure 1. Figure 1 also shows the spectrogram of the observed speech acquired from different irradiated objects, which are common everyday objects. From Figure 1, it can be seen that there are varying degrees of noise and a lack of speech components (see Figure 1c 2–5 kHz), and certain frequency bands are enhanced (see Figure 1c 5–8 kHz). We selected the observed speech in Figure 1b PET bottle as the processing object that has the greatest lack of speech components at the frequency band of 3–8 kHz with a noise of about 30 dB.



(**a**) (**b**) (**c**)

Figure 1. *Cont.*

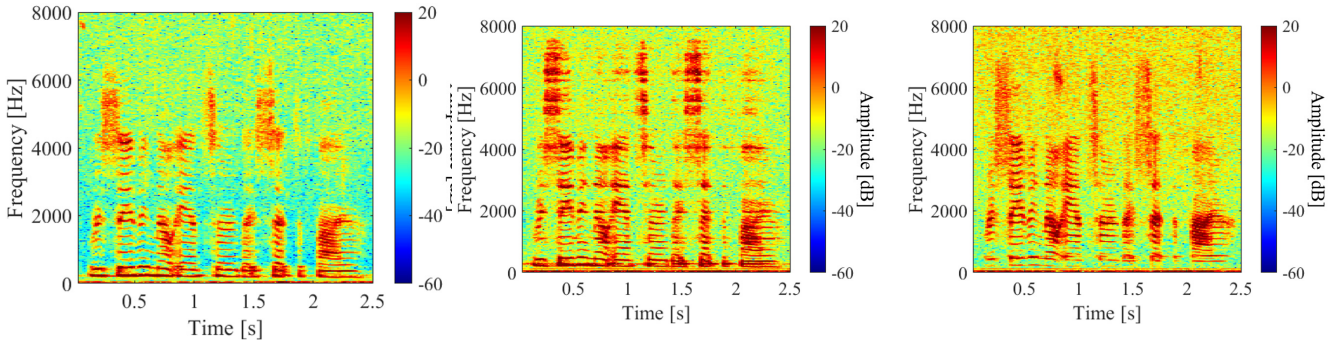


Figure 1. Power spectrum of observed speech acquired from different irradiated objects, i.e., (**a**) clean speech, (**b**) PET bottle, (**c**) cardboard box, (**d**) printing paper, (**e**) aluminum sheet, and (**f**) plastic plate. All objects were placed about 0.1 m from the sound source, and the output sound pressure was 85 dB(A).

2.2. Stft-Based Speech Enhancement

Figure 2 shows the power spectrum of clean speech (a), observed speech (b), amplitude response of the observed object (c), and phase difference between the observed speech and clean speech (d). As shown in Figure 2c, the amplitude response tends to decrease at high frequencies. From the left figures of Figure 2a,b, when the amplitude of the speech component is sufficiently smaller than the noise, the speech component in the frequency band of 4–8 kHz becomes unobservable. As shown in the right figures of Figure 2a,b, the phase spectrum has a complicated structure, so it is difficult to directly learn the mapping between the phase spectrum of observed speech and clean speech. As shown in Figure 2d, the phase difference between the observed speech and clean speech is strongly related to the amplitude of the observed speech. Therefore, with the STFT-based method, the power spectrum of enhanced speech is first estimated from the power spectrum of the observed speech, then the phase difference between the enhanced speech and observed speech is estimated using the power spectrum of the observed speech. Most speech components above 4 kHz are unvoiced, and the human ear is not sensitive to the phase of such unvoiced components, similar to white noise. Therefore, to increase the phase-restoration accuracy of the voiced components, only the 0–4 kHz frequency band is processed during phase restoration. For frequencies above 4 kHz, the phase of the observed speech is applied.

Figure 3 shows the processing procedure of the STFT-based method. Both the power and phase spectra are processed using DNNs. In the learning stage, the logarithmic power spectrum $|\boldsymbol{X}^{\mathrm{LPS}}| \in \mathbb{R}^{k\times m}$, $|\boldsymbol{Y}^{\mathrm{LPS}}| \in \mathbb{R}^{k\times m}$ (k, m are frequency and frame indices, respectively) and the phase spectrum $\boldsymbol{\phi}_x \in \mathbb{R}^{k\times m}$, $\boldsymbol{\phi}_y \in \mathbb{R}^{k\times m}$ are first extracted from clean speech $\boldsymbol{x}$ and observed speech $\boldsymbol{y}$ by Fourier transform. In the DNN for amplitude-spectrum reconstruction, the input and target of the network are $|\boldsymbol{Y}^{\mathrm{LPS}}|$ and $|\boldsymbol{X}^{\mathrm{LPS}}|$, respectively, and optimized by minimizing the mean squared error (MSE) between them. The structure of this DNN consists of two long short-term memory (LSTM) layers and three fully-connected layers, as shown in Figure 4. The activation function uses a rectified linear unit (ReLU). The input of the DNN for phase-spectrum estimation is $|\boldsymbol{Y}^{\mathrm{LPS}}|$ and the target is the phase difference $\boldsymbol{\phi}^{\mathrm{PD}} \in \mathbb{R}^{k\times m}$ obtained from $\boldsymbol{\phi}_x$ and $\boldsymbol{\phi}_y$. Since the phase spectrum has a period of 2π, the loss function used for learning is defined using Equation (1), i.e., the sum of the cosine distances between the input phase difference and target phase difference for each frame at a low-frequency band of 0–4 kHz.

$$\mathrm{Loss}_{\mathrm{PD}}(m) = \sum_{k=0}^{\frac{K}{4}+1} \left(1 - \cos\left(\boldsymbol{\phi}^{\mathrm{PD}}(k,m) - \hat{\boldsymbol{\phi}}^{\mathrm{PD}}(k,m)\right)\right) \tag{1}$$

As shown in Figure 5, the structure of the DNN for phase difference estimation consists of five convolutional layers. The activation function uses gated linear units (GLU) [18]. Both DNNs used Adam [19] to calculate the gradient.

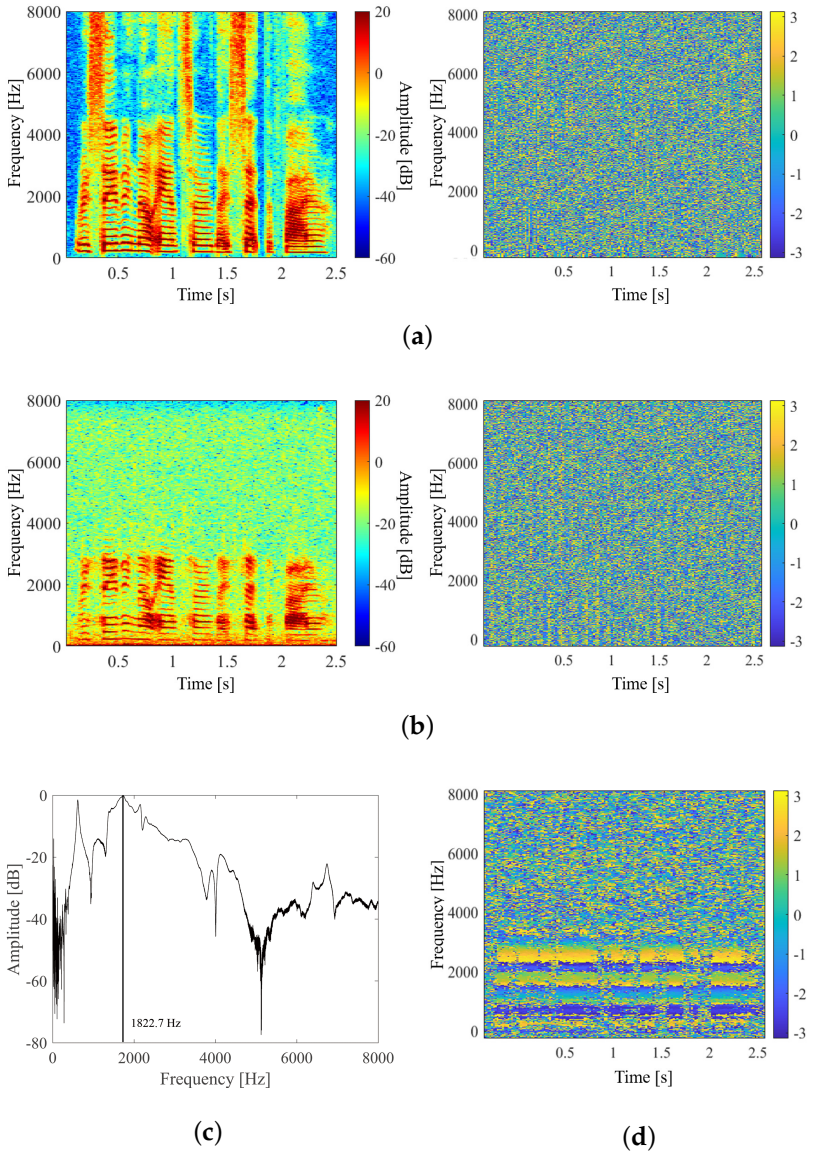


Figure 2. Power and phase spectrum of clean speech and observed speech and phase difference. (**a**) Power spectrum of clean speech (**left**) and its phase spectrum (**right**). (**b**) Power spectrum of observed speech (**left**) and its phase spectrum (**right**). (**c**,**d**) Amplitude response and phase difference between (**a**,**b**), respectively.

In the speech enhancement stage, the logarithmic power spectrum $|\boldsymbol{X}^{\mathrm{LPS}}|$ of the recorded speech is input to both DNNs, and the estimated logarithmic power spectrum $|\hat{\boldsymbol{Y}}^{\mathrm{LPS}}|$ and estimated phase difference $\hat{\boldsymbol{\phi}}^{\mathrm{PD}}$ are calculated. The reconstructed phase is obtained by adding the estimated phase difference $\hat{\boldsymbol{\phi}}^{\mathrm{PD}}$ and phase $\boldsymbol{\phi}^{\mathrm{PD}}$ of the observed speech. Finally, the enhanced speech can be obtained by inverse Fourier transform using $|\hat{\boldsymbol{Y}}^{\mathrm{LPS}}|$ and $\hat{\boldsymbol{\phi}}^{\mathrm{PD}}$.

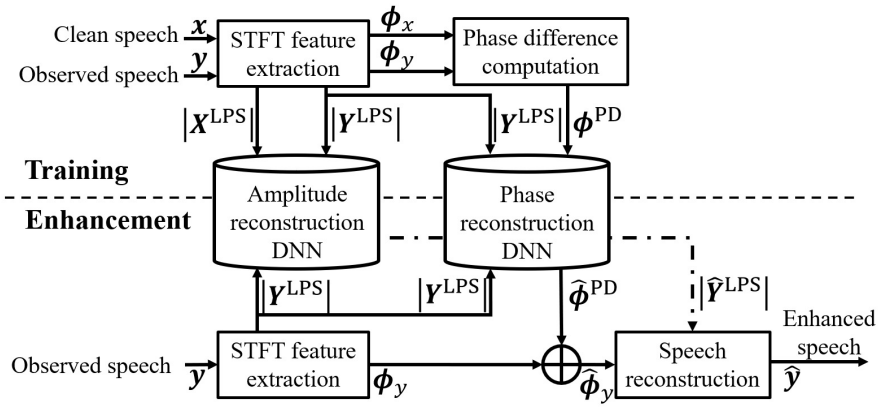


Figure 3. Block diagram of proposed STFT-based method.

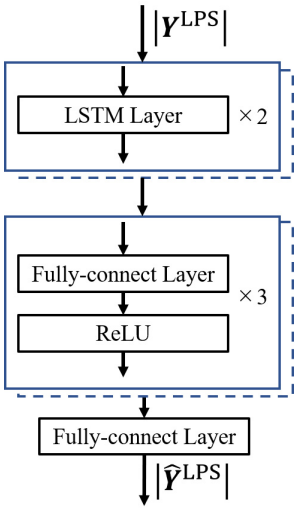


Figure 4. Structure of DNN for amplitude spectrum reconstruction.

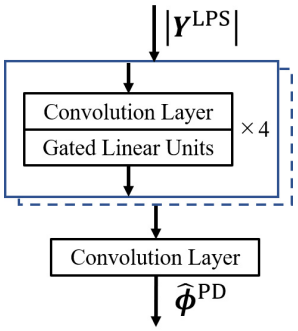


Figure 5. Structure of DNN for phase spectrum reconstruction.

2.3. Waveform-Based Speech Enhancement

Figure 6 shows the processing procedure of the waveform-based method. As described above, speech shows the obvious harmonic structure in the low-frequency bands and white noise distribution in the high-frequency bands. Therefore, with the waveform-based method, low-frequency and high-frequency components are processed separately.

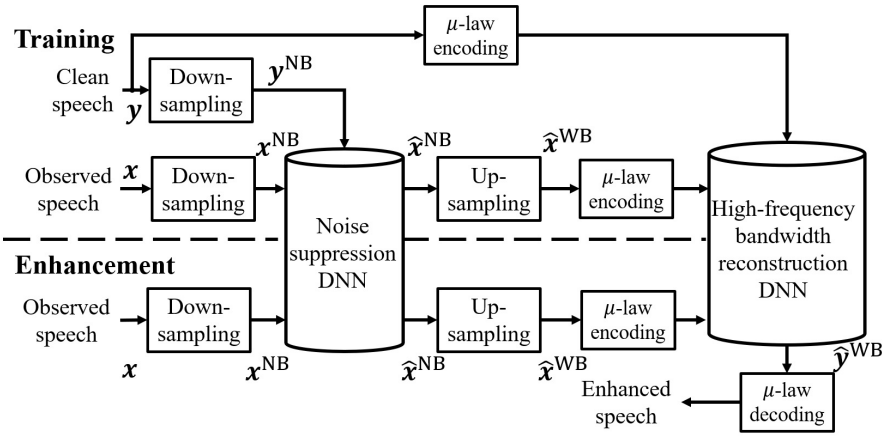


Figure 6. Block diagram of the proposed waveform-based method.

In the learning stage, since the high-frequency components of the observed speech $\boldsymbol{x}$ do not have the harmonic structure of speech, the high-frequency components of $\boldsymbol{x}$ and clean speech $\boldsymbol{y}$ are first removed by down-sampling to increase the accuracy of low-frequency component restoration. Down-sampled speech $\boldsymbol{x}^{\mathrm{NB}}$ and $\boldsymbol{y}^{\mathrm{NB}}$ are then input to the noise suppression DNN to train the network. Then, the low-frequency-enhanced speech $\hat{\boldsymbol{x}}^{\mathrm{NB}}$ are up-sampled to match the length of the clean speech. Finally, the up-sampled speech $\hat{\boldsymbol{x}}^{\mathrm{WB}}$ and $\boldsymbol{y}$ are each 8-bit quantized by μ-law [20] and used to train the DNN for high-frequency restoration.

The MSE is the most common loss function for deep learning. If set MSE as the loss function, its gradient is given by the derivative of MSE that $\hat{\boldsymbol{x}}^{\mathrm{NB}} - \boldsymbol{y}^{\mathrm{NB}}$. Here, the power of the speech signal concentrates in the low-frequency range. In other words, $\hat{\boldsymbol{x}}^{\mathrm{NB}} - \boldsymbol{y}^{\mathrm{NB}}$ is almost determined by the amplitude difference in the low-frequency bands, which is beneficial to the restoration of low-frequency signals. Moreover, when training a network on the basis of the MSE, the waveform of the recovered high-frequency component may be smoothed more than clean speech [21]. Therefore, the network with the MSE as the loss function is used for low-frequency band noise suppression. As shown in Figure 7, the noise-suppression DNN consists of convolutional layers with a dilated convolution structure of eight layers. The loss function and activation functions use the MSE and parametric ReLu (PReLU) [22], respectively.

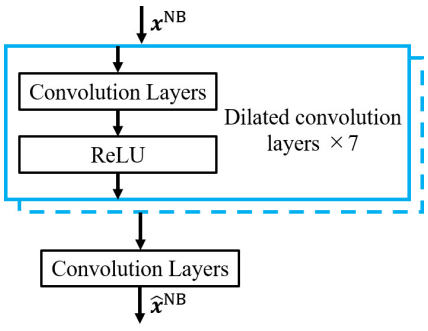


Figure 7. Structure of noise suppression DNN.

In the high-frequency component restoration process, recurrent neural networks (RNN) are used to predict the sample at time n by using the previous samples. By 8-bit quantization of the waveform, the amplitude of the waveform becomes an integer of 0–255, which will be input to the network. The output of the last fully connected layer is the probability distribution of each sample, which is a 256-dimensional vector. Figure 8 shows the network structure. The DNN for high-frequency component reconstruction consists of

two LSTM layers and two fully-connected layers and processes the input waveform sample by sample. Equation (2) shows the processing of the LSTM layer at time n.

$$\boldsymbol{S}(n) = \mathcal{G}(\boldsymbol{S}(n-1), \hat{x}^{\mathrm{WB}}(n)) \tag{2}$$

Here, $\boldsymbol{S}(n)$ is the LSTM output, $\hat{x}^{\mathrm{NB}}(n)$ is the network input, and $\mathcal{G}(\cdot)$ is the activation function of the LSTM. The output $\hat{y}^{\mathrm{WB}}(n)$ at time step n is based on the conditional probability of Equation (3).

$$p(\hat{y}^{\mathrm{WB}}(n)|\hat{x}^{\mathrm{WB}}(0), \hat{x}^{\mathrm{WB}}(1), \hat{x}^{\mathrm{WB}}(2), ..., \hat{x}^{\mathrm{WB}}(n)) = \mathrm{FC}(\boldsymbol{S}(n)) \tag{3}$$

Here, $\mathrm{FC}(\cdot)$ is the output of the fully-connected layer. The DNN for high-frequency restore is a classification network with cross-entropy as the loss function.

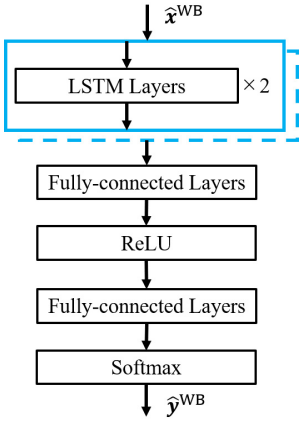


Figure 8. Structure of DNN for high-frequency restoration.

In the speech enhancement stage, high-frequency components are first removed from the observed speech $\boldsymbol{x}$ by down-sampling. The down-sampled speech $\boldsymbol{x}^{\mathrm{NB}}$ is then subjected to noise suppression by using the noise suppression DNN to enhance the low-frequency components. Then, up-sample $\hat{\boldsymbol{x}}^{\mathrm{NB}}$, which is the result of low-frequency enhancement and 8-bit quantized with μ-law. The results are input to the DNN for high-frequency component restoration, and the high-frequency components are reconstructed. Finally, the restored speech is obtained by decoding the output $\hat{y}^{\mathrm{WB}}$ of the DNN for high-frequency restoration using μ-law.

3. Evaluation Experiments

This section introduces the data collection for training and the hyperparameter setting of the network for objective experiments to evaluate the effectiveness of the proposed methods.

3.1. Training Data Set-Up

Speech data for network training were recorded using an LDV. The recording conditions are listed in Table 2, The experimental setup is shown in Figure 9, and the recording landscape is shown in Figure 10. An empty PET bottle with a capacity of 0.5 L was used as the measurement object. By using an empty PET bottle, the effect of reverberation due to the air inside the bottle can be considered. We set 4620 sentences in TIMIT Acoustic Phonetic [23] as the training database. Each sentence was recorded twice to create the experimental data using a total of 9240 sentences. Of these, 9000 (about 8 h) were used for training the network, and 240 (about 12 min) were used for evaluation.

Table 2. Experimental conditions.

Environment	**Sound-Proof Room**
Ambient noise level	20.8 dB
Sampling frequency	16 kHz
Sound pressure of sound source	85 dB(A)
Quantization bit rate	16 bits
Data	TIMIT Acoustic Phonetic Continuous Speech Corpus 9000 files (8 h) for training 240 files (12 min) for validation
Audio interface	Roland OCTA-CAPTURE UA-1010
LDV Equipment type	Polytec VibroFlex Xtra VFX-I-120
Software & libraries	Matlab R2022a, Deep learning HDL tool box v1.3
Vibrating object	PET-bottle

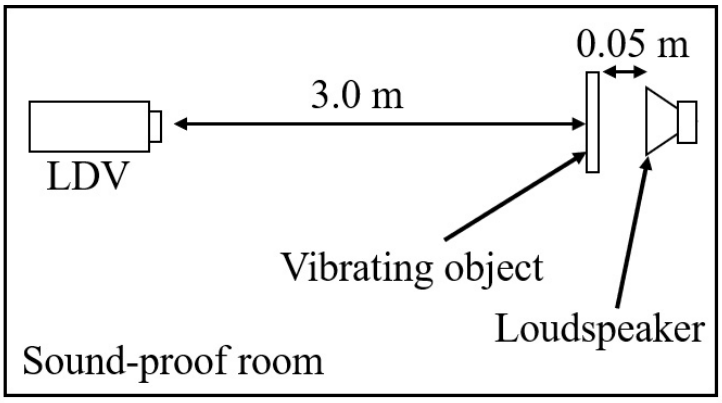


Figure 9. Experimental setup.

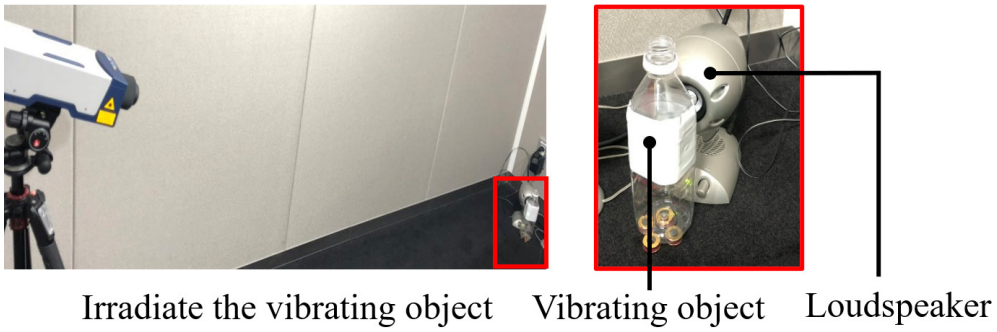


Figure 10. Experimental recording landscape.

3.2. Hyperparameters and Training Conditions

3.2.1. Stft-Based Method

When extracting STFT features, the Fourier transform length was set to 1024, and the frame shift length was set to 256. Considering the symmetry of the discrete Fourier transform, the input and output dimensions of the DNN for amplitude-spectrum reconstruction were set to 513 for each frame. The number of units in the LSTM layer and intermediate fully-connected layer was set to 1024, and the learning rate was set to 0.001. In the phase-recovery DNN, considering the continuity of the spectrum in the frame direction, a total of five frames were input, including two frames adjacent before and after the current

one-output frame. In the convolutional layers, the convolution kernel size of the first layer was set to 5×9, and the others were set to 1×9. The number of kernels in the convolutional layer was 1 in the output layer and 128 in other layers. The learning rate was set to 0.00001.

3.2.2. Waveform-Based Method

The input and output of the noise-suppression DNN were waveforms of 2048 samples. The DNN consists of eight convolutional layers with a dilated convolution structure, and the dilated factors are set to $2^n, n = 0, 1, \ldots, 7$. The convolution kernel size was 9×1. The number of kernels in the output layer was 1, and the other layers were 128. The learning rate was set to 0.0001. The number of units of the LSTM layer and fully-connected layer used for the high-pass reconstruction DNN was set to 1024. When training the network, the current time step sample was predicted using the previous 480 samples with backpropagation through time (BPTT) [24]. The learning rate was set to 0.0001, the same as the noise-suppression DNN.

3.3. Evaluation Experiments for Stft-Based Method

To confirm the effectiveness of the phase spectrum with the proposed STFT-based method, we conducted an evaluation experiment using the observed speech, enhanced speech restored only by amplitude-spectrum reconstruction, enhanced speech restored by both amplitude spectrum and phase spectrum constructed using the GLA [25,26] (200 iterations) with the initial phase of observed speech, and enhanced speech with the proposed STFT-based method. The amplitude spectrogram of each speech is shown in Figure 11. It can be seen from Figure 11 that the result Figure 11e of the proposed STFT-based method has less noise than the result Figure 11c,d. And in particular, it can be confirmed that the noise in the band below 0–1 kHz was suppressed, and the harmonic structure became clear.

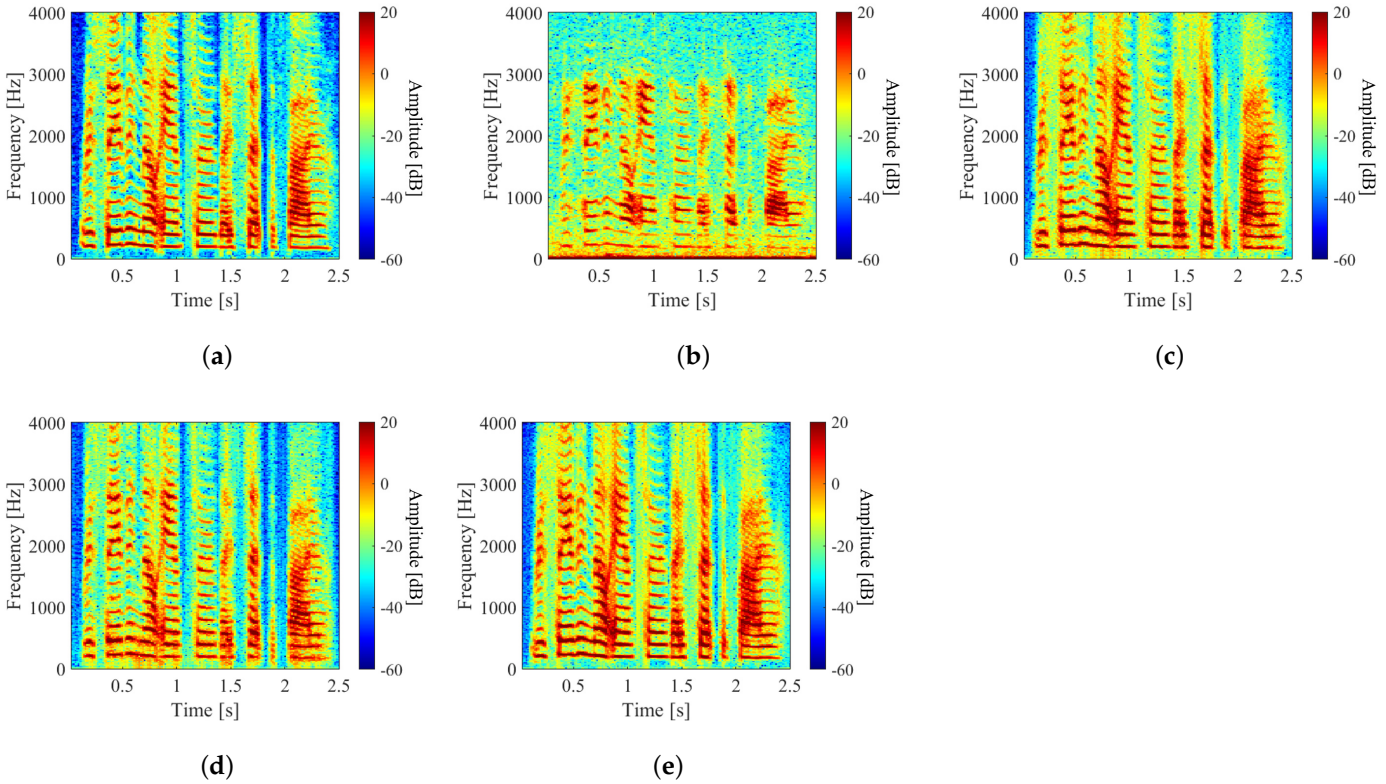


Figure 11. Spectrogram of speech-enhancement results: (**a**–**e**) are clean speech, observed speech, enhanced speech with reconstructed amplitude and observed phase, enhanced speech with reconstructed amplitude and phase reconstructed using GLA, and enhanced speech with proposed STFT-based method], respectively.

To evaluate the proposed phase reconstruction method, we conducted an evaluation experiment using the phase of the observed speech, the phase applied to the GLA, and the phase reconstructed with the proposed method. Figure 12 shows the cosine distance between the restored phase and clean speech phase, and the smaller the cosine distance, the higher the accuracy of phase reconstruction. The cosine distance with the proposed method is about 0.4 smaller than that of using the GLA in the frequency band of 0–4 kHz. In the 0–8 kHz band, the cosine distance with the STFT-based method was about 0.2 smaller than that of the GLA. These results indicate the effectiveness of the STFT-based method for phase reconstruction.

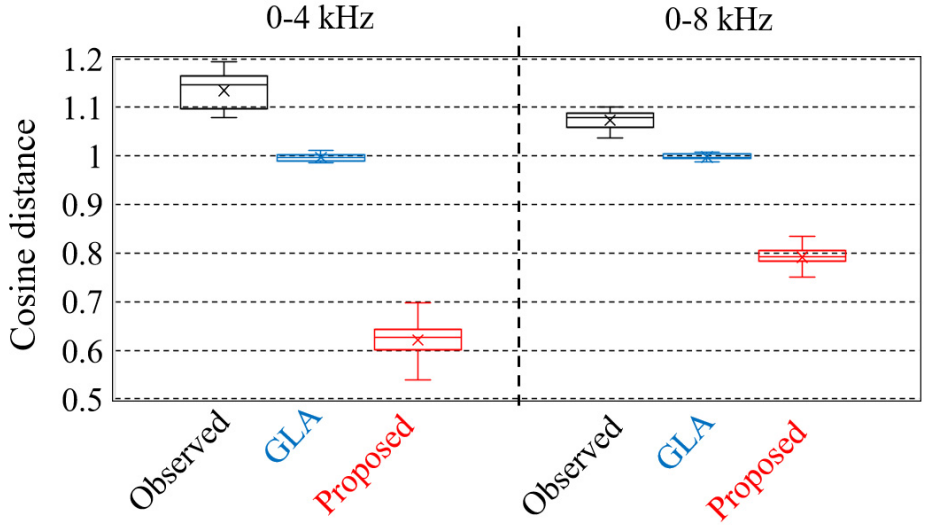


Figure 12. Cosine distance of reconstructed phases with each method.

3.4. Evaluation Experiment for Waveform-Based Method

To confirm the effectiveness of the proposed waveform-based method, we conducted evaluation experiments using the observed speech, speech with only noise suppression, and that with the waveform-based method. Figure 13 shows the amplitude spectrum of each speech. The high-frequency components were restored with the waveform-based method and the noise in high-frequency bands was smaller than those in Figure 13c. As shown in Figure 13c, the noise-suppression DNN of the waveform-based method sufficiently enhanced speech in the low-frequency band. Figure 13b–d were evaluated in each frequency band using the criterion of log-spectral distance (LSD) [27]. The LSD is defined by the following equation, and the smaller the value, the less speech degradation.

$$\mathrm{LSD} = \sqrt{\frac{1}{f_1 - f_0} \sum_{k=f_0}^{f_1} \left(10 \log_{10} \frac{P(k)}{\hat{P}(k)} \right)^2} \tag{4}$$

Here, f_0 and f_1 are the frequency band ranges, and P and $\hat{P}$ are the clean and estimated speech powers, respectively. Figure 14 shows the LSD in each frequency band. The LSD was about 0.3 dB smaller in the 4–8 kHz band where the high-frequency band was restored. In the 0–4 kHz band, the difference in the LSD between the presence and absence of high-frequency restoration was very small. Therefore, the effectiveness of the waveform-based method was confirmed.

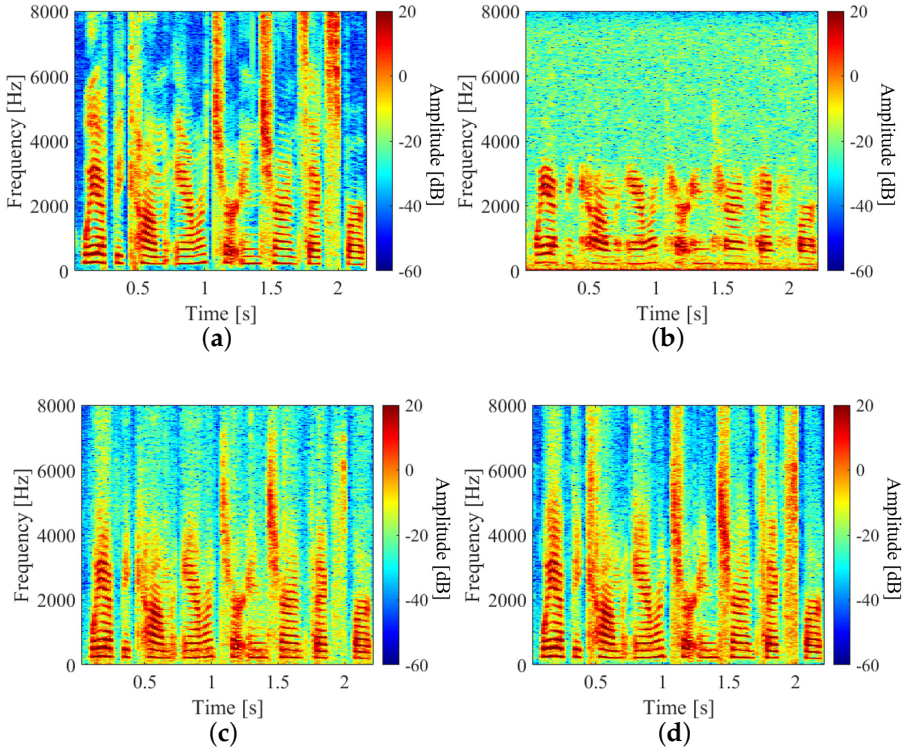


Figure 13. Spectrogram of speech-enhancement results: (**a**–**d**) are clean speech, observed speech, speech obtained by only noise suppression, speech enhanced with proposed waveform-based method

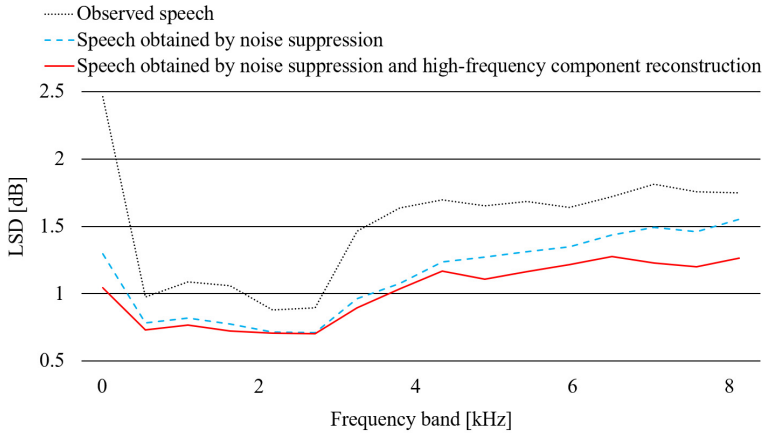


Figure 14. LSD of enhanced speech with each method at each frequency band.

3.5. Evaluation Results and Discussion

To confirm the effectiveness of the proposed methods, we conducted speech enhancement experiments using speech recorded with an LDV. We used wide-band perceptual evaluation of speech quality (PESQ), LSD, and short-time objective intelligibility (STOI) as evaluation criteria to compare sound quality, degree of deterioration, and intelligibility. The higher the PESQ and STOI, the higher the sound quality and intelligibility. The objects for comparison were clean speech, observed speech, the result of speech enhancement based on the conventional method [9], the result of the STFT-based method, and the result of the waveform-based method. Figure 15 shows the spectrogram, and Table 3 shows the evaluation results of each speech. From Figure 15, the noise was suppressed with both proposed methods. With the conventional method, band-pass filters were used to completely remove speech components that were lacking due to the frequency response, and the proposed methods restored these components. As shown in Table 3, it was confirmed that the PESQ

scores of the proposed methods were almost the same as that of the conventional method; the PESQ was about 0.5 higher than the observed speech. The proposed methods had smaller LSDs than the conventional method. With the conventional method, the LSD was larger than the proposed methods because it is difficult to enhance speech in the 4–8 kHz frequency band of the observed speech where the noise is larger than the speech. For STOI, the proposed methods improved by about 0.08 compared with the recorded speech, and improved by about 0.06 compared with the conventional method.

It can be concluded from the above experimental data that, compared with the conventional methods, the proposed methods significantly improve the sound quality of the observed speech. Moreover, compared with the single-stage DNN-based method, the two-stage processing can also better deal with the various deterioration in the observed speech.

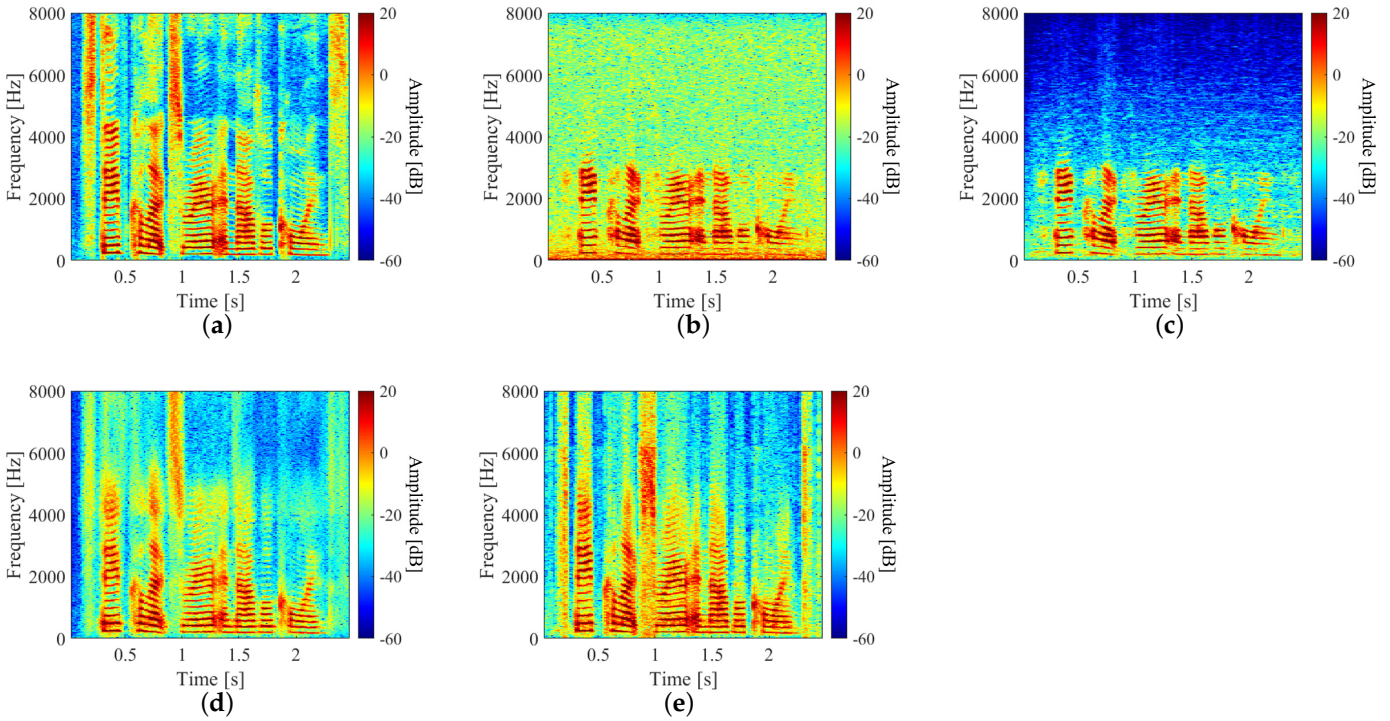


Figure 15. Examples of clean, observed, and enhanced speech spectrograms: (**a**–**e**) are clean speech, observed speech, speech enhanced with the conventional method, speech enhanced with proposed STFT-based method, and speech enhanced with the proposed waveform-based method, respectively.

Table 3. Objective evaluation result of each method.

	PESQ Score	LSD [dB]	STOI Score
(b)	1.76 ± 0.40	1.62 ± 0.10	0.85 ± 0.04
(c)	2.25 ± 0.35	2.17 ± 0.30	0.87 ± 0.04
(d)	2.25 ± 0.30	1.09 ± 0.08	0.93 ± 0.02
(e)	2.35 ± 0.30	1.11 ± 0.08	0.94 ± 0.03

(b) observed speech, (c) conventional method, (d) proposed method 1, (e) proposed method 2, respectively.

4. Conclusions

We proposed two speech-enhancement methods with two-stage processing based on DNNs for reducing distortion of speech observed with an LDV, i.e., a method using two-stage processing with amplitude-spectrum reconstruction and phase-spectrum estimation DNNs (STFT-based method), and a method with two DNNs of noise suppression and high-frequency component reconstruction (waveform-based method). Since an LDV uses a laser to measure vibrations caused by sound waves, various types of deterioration are mixed in due to the frequency response of the irradiated object. Due to these various types of deterioration, there is a problem that the accuracy of the conventional speech

enhancement method for LDV is unsatisfactory. Objective evaluation experiments showed that the proposed methods improved enhancement accuracy by comparing them with the conventional method in terms of sound quality, degree of deterioration, and intelligibility. The methods solve the problem of sound-quality degradation when using LDVs to acquire speech. These methods can be applied to recording sounds from areas where people cannot enter, such as disaster sites, or to recording target speech in noisy environments, such as recording the sound of a stage at a concert.

We plan to study differences in enhancement accuracy due to the loss function in the proposed STFT-based method. We also plan to investigate a speech enhancement method that takes into account the deterioration of the observed speech due to changes in the material of the measurement object and an increase in the distance between the irradiated object and LDV.

Author Contributions: C.C and T.N. conceived the proposed method. C.C. developed the method and conducted the experiments. C.C wrote this manuscript and modified the figures and expressions of the equations. T.N. is an administrator of this study. K.I. and T.N. supervised this research. All authors discussed the results and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported in part by the Ritsumeikan Advanced Research Academy (RARA), Ritsumeikan Global Innovation Research Organization (R-GIRO), and by JSPS KAKENHI Grant Nos. JP19H04142, JP21H03488, JP21H04427 and JP21K18372.

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. Clark, M.A. An acoustic lens as a directional microphone. *Trans. IRE Prof. Group Audio* **1953**, *25*, 1152–1153.
2. Taylor, K.J. Absolute measurement of acoustic particle velocity. *J. Acoust. Soc. Am.* **1976**, *59*, 691–694. [CrossRef]
3. Shang, J.H.; He, Y.; Liu, D.; Zang, H.G.; Chen, W.B. Laser Doppler vibrometer for real-time speech-signal acquirement. *Chin. Opt. Lett.* **2009**, *7*, 732–733. [CrossRef]
4. Leclère, Q.; Laulagnet, B. Nearfield acoustic holography using a laser vibrometer and a light membrane. *J. Acoust. Soc. Am.* **2009**, *126*, 1245–1249. [CrossRef] [PubMed]
5. Avargel, Y.; Cohen, I. Speech measurements using a laser Doppler vibrometer sensor: Application to speech enhancement. In Proceedings of the 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, Edinburgh, UK, 30 May 2011; pp. 109–114. [CrossRef]
6. Malekjafarian, A.; Martinez, D.; Brien, E.J.O. The feasibility of using laser Doppler vibrometer measurements from a passing vehicle for bridge damage detection. *Shock Vib.* **2018**, *2018*, 1–10. [CrossRef]
7. Chen, D.M.; Xu, Y.F.; Zhu, W.D. Identification of damage in plates using full-field measurement with a continuously scanning laser Doppler vibrometer system. *J. Sound Vib.* **2018**, *422*, 542–567. [CrossRef]
8. Aygün, H.; Apolskis, A. The quality and reliability of the mechanical stethoscopes and Laser Doppler Vibrometer (LDV) to record tracheal sounds. *Appl. Acoust.* **2020**, *161*, 1–9. [CrossRef]
9. Li, W.H.; Liu, M.; Zhu, Z.G.; Huang, T.S. LDV remote voice acquisition and enhancement. In Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, 20–24 August 2006; pp. 262–265.
10. Peng, R.H.; Xu, B.B.; Li, G.T.; Zheng, C.S.; Li, X.D. Long-range speech acquirement and enhancement with dual-point laser Doppler vibrometers. In Proceedings of the IEEE 23rd International Conference on Digital Signal Processing, Shanghai, China, 19–21 November 2018; pp. 1–5.
11. Xie, Z.; Du, J.; McLoughlin, I.; Xu, Y.; Ma, F.; Wang, H. Deep neural network for robust speech recognition with auxiliary features from laser-Doppler vibrometer sensor. In Proceedings of the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 17–20 October 2016; pp. 1–5. [CrossRef]
12. Lü, T.; Guo, J.; Zhang, H.Y.; Yan, C.H.; Wang, C.J. Acquirement and enhancement of remote speech signals. *Optoelectron. Lett.* **2017**, *13*, 275–278. [CrossRef]
13. Li, K.H.; Lee, C.H. A deep neural network approach to speech bandwidth expansion. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, 19–24 April 2015; pp. 4395–4399.

14. Lotter, T.; Vary, P. Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-Gaussian speech modelling. In Proceedings of the 12th European Signal Processing Conference, Vienna, Austria, 6–10 September 2004; pp. 1457–1460.
15. Rethage, D.; Pons, J.; Serra, X. A Wavenet for speech denoising. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 5069–5073.
16. Krawczyk, M.; Gerkmann, T. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE ACM Trans. Audio Speech Lang. Process.* **2018**, *22*, 1931–1940. [CrossRef]
17. Takamichi, S.; Saito, Y.; Takamune, N.; Kitamura, D.; Saruwatari, H. Phase reconstruction from amplitude spectrograms based on von-mises-distribution deep neural Network. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 286–290
18. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the 34th International Conference on Machine Learning, Ningbo, China, 6–11 August 2017; pp. 933–941.
19. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
20. Recommendation, I.G. 711: Pulse code modulation (PCM) of voice frequencies. *Int. Telecommun. Union* **1988**.
21. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv* **2017**, arXiv:1609.04802.
22. He, K.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1024–1034.
23. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L. *Acoustic-Phonetic Continuous Speech Corpus CD-ROM NIST Speech Disc 1-1.1*; NASA STI/Recon Tech. Rep. LDC93S1; Linguistic Data Consortium: Philadelphia, PA, USA, 1993; Volume 93.
24. Werbos, P.J. Backpropagation through time: What it does and how to do it. *Proc. IEEE* **1990**, *78*, 1550–1560. [CrossRef]
25. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Audio Speech Lang. Process.* **1984**, *32*, 236–243. [CrossRef]
26. Perraudin, N.; Balazs, P.; Sondergaard, P.L. A fast Griffin-Lim algorithm. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
27. Wang, P.; Wang, Y.; Liu, H.; Sheng, Y.; Wang, X.; Wei, Z. Speech enhancement based on auditory masking properties and log-spectral distance. In Proceedings of the 3rd International Conference on Computer Science and Network Technology, Dalian, China, 12–13 October 2013; pp. 1060–1064. [CrossRef]