Speech Enhancement for Optical Laser Microphone With Deep Neural Network

Chengkai Cai* Kenta Iwai* Takanobu Nishiura* and Yoichi Yamashita*
* Ritsumeikan University, Shiga, Japan
E-mail: gr0370hv@ed.ritsumei.ac.jp, iwai18sp@fc.ritsumei.ac.jp, nishiura@is.ritsumei.ac.jp, yyama@is.ritsumei.ac.jp

***Abstract*—The development of distant-talking speech measurement systems is drawing attention since it can be used for security and disaster relief. An optical laser microphone that measures the vibration caused by sound using laser beam is suitable for the system. It can capture only the target sound even in a noisy place. However, the sound acquired by the optical laser microphone depends on the irradiated objects, and most of the objects will cause degradation of sound quality. Due to various deteriorations in the sound acquired by the laser microphone, the results of the conventional speech enhancement methods are insufficient. Therefore, in this paper, we propose a speech enhancement method using a two-stage process of noise suppression and acoustic structure reconstruction for optical laser microphones and conduct an objective experiment to evaluate the effectiveness of the proposed method.**

I. Introduction

Microphones play an important role in daily life, and a variety of microphones have been developed to meet different needs. Ordinary microphones obtain sound signals by converting sound pressure into electrical signals through an internal diaphragm. However, as the distance between the microphone and sound source increases, the power of the sound waves is attenuated when the sound propagates in the air, and therefore it is difficult to acquire distant sounds with ordinary microphones. Since a distant-talking speech measurement system is required for crime prevention security and disaster relief, parabolic microphones and shotgun microphones were developed to obtain sounds at a distance[1]. These microphones acquire the distant sounds by changing the shape of the receiving part and the position of the diaphragm. These kinds of microphones receive all sounds regardless of a target sound source at a distance or the noise around the microphone, which makes them unusable in a noisy environment. Therefore, an acoustic measurement system called an optical laser microphone was developed that measures vibration caused by a sound using a laser beam. Since the optical laser microphone directly measures the vibration of the object near the target sound source, it will not be affected by the noise around the microphone.

However, the quality of the sound acquired by the optical laser microphone usually deteriorates because of the following two reasons. The first is the change in reflected light intensity due to the rough surface of the irradiated object, causing a noise mix in the signal, and the second is the objects cannot vibrate at high frequency, causing a lack of high-frequency information of the observed signal. Therefore, to improve the sound quality, there are two main tasks in the processing of the observation signal, one is noise reduction and the other is high-frequency speech component reconstruction.

In recent years, deep learning has been widely used in various fields of signal processing and has proven its effectiveness in acoustic signal processing, such as speech generation [2], voice conversion [3][4], and speech enhancement [5][6]. The speech is time-series data and is usually processed in the frequency domain using a Fourier transform. Recently, processing based on the sound waveforms has also been proposed, such as the use of recurrent neural networks [7][8] and WaveNet [9][10]. However, it is difficult to directly use these methods because the observed sound of the optical laser microphone has various distortions.

In this paper, we propose a speech enhancement method for observed speech acquired by an optical laser microphone. The proposed method uses a waveform-based deep neural network (DNN) and consists of noise reduction and high-frequency component reconstruction. To confirm the effectiveness of the proposed method, we perform objective evaluation experiments for the speech measured by the optical laser microphone and evaluate the sound quality of the speech before and after processing.

II. Acoustic Measurement using Optical Laser Microphone

After sound is generated, the vibration caused by the sound will be transmitted to the nearby objects through the air. When irradiating a laser beam on the object, the amplitude and phase of the reflected laser will be changed due to the vibration. By focusing on the amplitude of the reflected laser, the vibration can be measured by using a photo-diode. Figure 1 shows a schematic diagram of the measurement by photo-diode, and the relationship between the acoustic signal and amplitude of measured reflected laser can be expressed as

$$S_r(t) = A_c(1 + s(t))\cos(2\pi F_c t + \phi_c), \tag{1}$$

where A_c, F_c, and ϕ_c are the amplitude, frequency, and phase of the reflected laser, respectively, and t is the time index.

Also, by focusing on the phase of reflected laser, the vibration can be measured by a laser Doppler vibrometer (LDV) that utilizes the Doppler effect of light. Figure 2 shows a schematic diagram of the measurement by LDV. The Doppler effect occurs when the laser light is reflected on the surface of the vibrating object. The frequency difference $f_{\mathrm{D}}(t)$

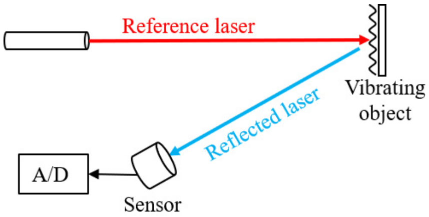


Fig. 1. Acoustic measurement by photo-diode.

between the reference laser and the reflected laser can be detected by using the interferometer. The relationship between the frequency difference $f_{\mathrm{D}}(t)$ and the vibration velocity is expressed as

$$f_{\mathrm{D}}(t) = \frac{2v(t)}{\lambda_0} = \frac{2}{\lambda_0} \cdot \frac{\mathrm{d}L_1(t)}{\mathrm{d}t}, \quad (2)$$

where λ_0 is the wavelength of the reference laser, $v(t)$ is the velocity of the vibration, and $L_1(t)$ is the optical path length from the laser transmitter to the detector in the LDV. Interference occurs when the reflected light and the reference light are superposed. The vibration velocity of the object can be measured by calculating the interval of the interference fringes that accompany changes in the optical path length. However, the vibration direction cannot be discriminated if only the interval of the interference fringes is used. Therefore, the frequency of the reference light is changed by the frequency shifter to determine the sign of the vibration velocity. The luminous intensity can be expressed as

$$I(t) = I_1 + I_2 + 2\sqrt{I_1 I_2 \mathrm{cos}(\rho)}, \quad (3)$$

$$\rho = 2\pi \frac{(L_2 - L_1(t))(\Delta f - f_D(t))}{c}, \quad (4)$$

where I_1 is the intensity of the reflected laser, I_2 is the intensity of the reference laser, L_2 is the optical path length of the reference light, Δf is the amount of frequency shift given by the frequency shifter, and c represents the velocity of light.

Since the laser beam has high directivity, which means that the component light waves travel together in a straight line without spreading apart, it is possible to measure the vibration generated at a distance. Also, since only the sound generated around the irradiated object is measured, it is not affected by the noise around the optical laser microphone. Because of the features above, the optical laser microphone is useful when receiving distant-talking speech. However, since the acoustic information is acquired through the vibrating object, the quality of the observed sound depends on the shape and vibration characteristics of the irradiated object. For example, when an object with a rough surface moves in a direction different from the measurement direction, $f_{\mathrm{D}}(t)$ changes not only due to $v(t)$ but also due to the unevenness of the object surface. In addition, the intensity of the reflected laser light decreases and stationary noise occurs. Therefore, this paper proposes a deep-learning-based speech enhancement method for the sound acquired by the optical laser microphone.

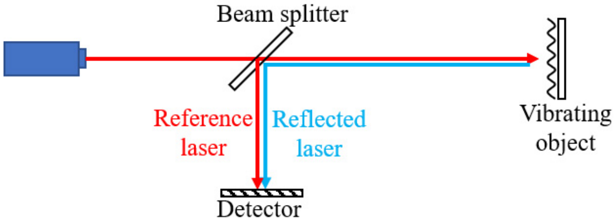


Fig. 2. Acoustic measurement by LDV.

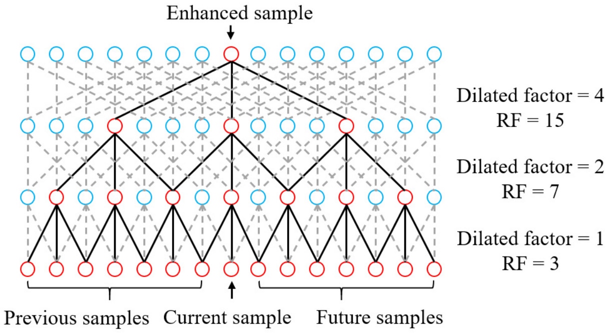


Fig. 3. Structures of dilated convolution.

III. Conventional Speech Enhancement Method Based on DNN

Conventional speech enhancement methods based on deep learning mostly convert time waveforms to a frequency spectrum. First, the spectrum obtained by applying a Fourier transform is separated into the power and phase components. Then, the adjusted power spectrum can be calculated by the trained DNN model. Finally, the waveform is reconstructed by the inverse Fourier transform using the adjusted power spectrum and unprocessed phase spectrum. However, since the high-frequency component of the observed signal is missing, if the phase spectrum of the deteriorated sound is used when restoring the waveform, artifacts in the high-frequency band will occur and the sound quality will be insufficient.

Wavenet is a typical network based on waveform processing that uses the dilated convolution shown in Fig. 3 to obtain a larger receptive field so that it is able to handle the time series. The skip-connection structure is used to accelerate the convergence. However, because of the various kinds of deterioration, the accuracy of the results estimated by directly modeling the waveform decreases.

Therefore, in this paper, a speech enhancement method with multi-stage processing for the speech observed by the optical laser microphone that has various distortions is proposed.

IV. Speech Enhancement for Optical Laser Microphones

In this section, a speech enhancement method based on waveform processing for the observed speech by the optical laser microphone is proposed. The experimental results described in the subsection C of section V shows that the CNN trained with mean square error (MSE) as the loss function has a good performance in the low-frequency band components of speech, yet it is not good in the high-frequency band components (see Fig. 10 (d)). However, the reconstruction of the high-frequency components depends on the accurate low-frequency components because the high-frequency components of the signal acquired by LDV are lacking. Since a single network is difficult to handle the various kinds of distortion, the proposed method consists of a noise suppression function in the low-frequency band and a reconstruction function in the high-frequency band. The processing diagram of the proposed method is shown in Fig. 4.

A. Noise suppression process

Since there are no speech-specific structures in the high-frequency component of the observed signal $\boldsymbol{x}$ (see Fig. 10 (a) and (b)), the high-frequency components are first removed by downsampling to reduce the impact of the noise at high frequency on the predict result. After that, the low-frequency component $\boldsymbol{x}^{\mathrm{NB}}$ is processed with the noise suppression DNN. Then, the noise suppressed signal $\hat{\boldsymbol{x}}^{\mathrm{NB}}$ is upsampled to match the length of the observed signal and quantized by μ-law [11]. The wideband speech $\hat{\boldsymbol{y}}^{\mathrm{WB}}$ is estimated by the high-frequency bandwidth reconstruction DNN, and the high-frequency components are extracted by the high-pass filter. Finally, the enhanced speech is obtained by adding the estimated high-frequency components to the processed low-frequency components.

In the noise suppression processing, the DNN is composed of $l+1$ convolutional layers and activation function rectified linear unit (ReLU) as shown in Fig. 5. First, the input feature maps are extracted from the waveform of the low-frequency component by

$$\boldsymbol{F}_{(1)} = \max(0, \boldsymbol{W}_{(1)} * \boldsymbol{x}^{\mathrm{NB}} + \boldsymbol{B}_{(1)}), \tag{5}$$

where $\boldsymbol{W}$ is the weights of the convolution kernel, $\boldsymbol{B}$ is the bias, and the subscript 1 means the first layer of $l+1$ layers. The middle layer is consisted of stacked convolutional layers and calculates the mapping relationships between the $\boldsymbol{F}_{(1)}$ and the reconstruction feature maps $\boldsymbol{F}_{(l)}$ by (6).

$$\boldsymbol{F}_{(m)} = \max(0, \boldsymbol{W}_{(m)} * \boldsymbol{F}_{(m-1)} + \boldsymbol{B}_{(m)}). \tag{6}$$

Here, m is the layers index and $m = [2, 3, ..., l]$. The output layer restores the output of last middle layer $\boldsymbol{F}_{(l)}$ to the waveform $\hat{\boldsymbol{x}}^{\mathrm{NB}} = \{\hat{x}^{\mathrm{NB}}(0), ..., \hat{x}^{\mathrm{NB}}(N-1)\}$ by

$$\hat{\boldsymbol{x}}^{\mathrm{NB}} = \boldsymbol{W}_{(l+1)} * \boldsymbol{F}_{(l)} + \boldsymbol{B}_{(l+1)}. \tag{7}$$

When the noise suppression network is optimized by minimizing the MSE as the loss function, the gradient is always updated along the direction of the minimum difference of amplitude between the observed signal and the target signal. Since the amplitude difference mainly depends on the low-frequency components, it is effective for the processing of low-frequency components. However, the structure of unvoiced sound in the high-frequency band has characteristics similar to noise, and it is difficult to extract features by the convolutional layer. Furthermore, it is difficult to estimate the details of high-frequency components when training the network with MSE [12]. Therefore, high-frequency components require additional processing, which is described in the next subsection.

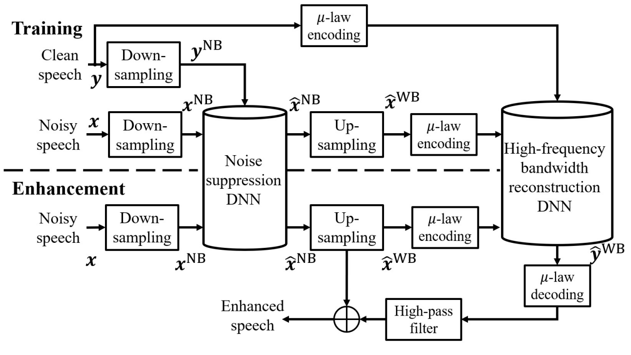

Training
Clean speech
$\boldsymbol{y}$
Down-sampling
$\boldsymbol{y}^{\mathrm{NB}}$
μ-law encoding
Noisy speech
$\boldsymbol{x}$
Down-sampling
$\boldsymbol{x}^{\mathrm{NB}}$
Noise suppression DNN
$\hat{\boldsymbol{x}}^{\mathrm{NB}}$
Up-sampling
$\hat{\boldsymbol{x}}^{\mathrm{WB}}$
μ-law encoding
High-frequency bandwidth reconstruction DNN
Enhancement
Noisy speech
Down-sampling
$\boldsymbol{x}$
$\boldsymbol{x}^{\mathrm{NB}}$
Up-sampling
$\hat{\boldsymbol{x}}^{\mathrm{NB}}$
μ-law encoding
$\hat{\boldsymbol{x}}^{\mathrm{WB}}$
$\hat{\boldsymbol{y}}^{\mathrm{WB}}$
μ-law decoding
Enhanced speech
High-pass filter

Fig. 4. Block diagram of waveform-based DNN.

B. High-frequency bandwidth reconstruction process

In the high-frequency component reconstruction process, the RNN is applied and cross entropy (CE) is chosen as the loss function. In this stage, the waveform processed by the noise suppression is first quantized into 8-bit integer values of 0-255 by applying μ-law. As shown in Fig. 6, the network is composed of two long short-term memory (LSTM) layers and two fully-connected layers. The processing of the LSTM layer at time step n is shown in

$$\boldsymbol{s}(n) = \mathcal{G}(\boldsymbol{s}(n-1), \hat{x}^{\mathrm{WB}}(n)), \tag{8}$$

where $\boldsymbol{s}$ is the output of the LSTM layer, $\hat{x}^{\mathrm{WB}}$ is the result of the noise suppression process, and $\mathcal{G}(\cdot)$ is the activation function. The activation function of the last fully-connected layer is softmax, which is used to output the probability of each label. The output $\hat{y}_t^{\mathrm{WB}}$ of the time step n is calculated by

$$p(\hat{y}^{\mathrm{WB}}(n)|\hat{x}^{\mathrm{WB}}(1), \hat{x}^{\mathrm{WB}}(2), ..., \hat{x}^{\mathrm{WB}}(n)) = \mathrm{FC}(\boldsymbol{s}(n)), \tag{9}$$

where $\mathrm{FC}(\cdot)$ is the output of the fully-connected layer. The observed speech $\hat{\boldsymbol{x}}^{\mathrm{WB}}$ obtained from the noise suppression process and the clean speech $\boldsymbol{y}$ is used to optimize the network.

This stage of processing uses information up to time step n to predict $\hat{y}^{\mathrm{WB}}(n)$ because the previous stage of processing effectively reduces the interference of noise on prediction results, thereby achieving a higher accuracy than conventional methods.

V. Evaluation experiment

In this section, we carried out a recording experiment using the optical laser microphone and compared the observed speech, and evaluated the effectiveness of the proposed method by an objective evaluation. The effectiveness of the proposed method was compared with that of the conventional methods [6] and [10].

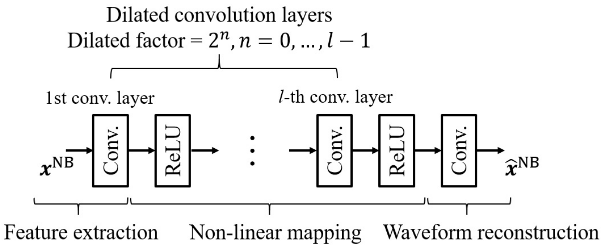

Dilated convolution layers
Dilated factor = $2^n, n = 0, \ldots, l-1$
1st conv. layer
l-th conv. layer
$\boldsymbol{x}^{\mathrm{NB}}$
Conv.
ReLU
Conv.
ReLU
Conv.
$\hat{\boldsymbol{x}}^{\mathrm{NB}}$
Feature extraction
Non-linear mapping
Waveform reconstruction

Fig. 5. Network for noise suppression.

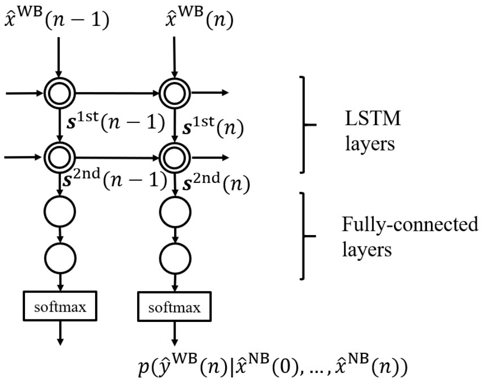

$\hat{x}^{\mathrm{WB}}(n-1)$
$\hat{x}^{\mathrm{WB}}(n)$
$\boldsymbol{s}^{1\mathrm{st}}(n-1)$
$\boldsymbol{s}^{1\mathrm{st}}(n)$
$\boldsymbol{s}^{2\mathrm{nd}}(n-1)$
$\boldsymbol{s}^{2\mathrm{nd}}(n)$
LSTM layers
Fully-connected layers
softmax
softmax
$p(\hat{y}^{\mathrm{WB}}(n)|\hat{x}^{\mathrm{NB}}(0), \ldots, \hat{x}^{\mathrm{NB}}(n))$

Fig. 6. Process of high-frequency band reconstruction at time step n.

A. Training data setup

The speech data for the training network was recorded using LDV. The experimental conditions are shown in Table I, the equipment arrangement is shown in Fig. 7, and the experimental landscape is shown in Fig. 8. An empty plastic bottle with a volume of 0.5 liters was used as the measurement object. As well as the material, the reverberation caused by the air inside the empty bottle could also affect the results of the recording experiment. 4620 speech files in the TIMIT Acoustic Phonetic corpus [14] were recorded twice as the experimental data. 9000 speech files (about 4 hours) were used for training the network, and 240 speech files (about 12 minutes) were used for evaluation.

B. Network setup

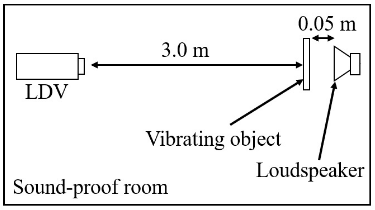


Fig. 7. Experimental arrangement.

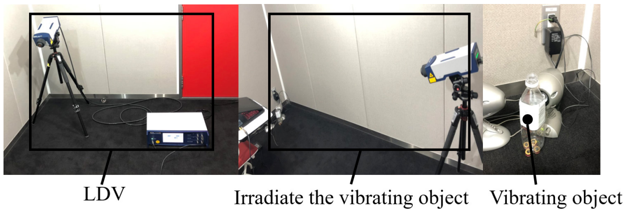


Fig. 8. Experimental landscape.

The network of noise suppression process is completely composed of convolutional layer. In order to determine the size of the convolution kernel, a comparison experiment as shown in Fig. 9 was conducted. Figure 9 shows the log spectral distance (LSD) for the noise suppression results obtained by the noise suppression DNN trained with different convolution kernel sizes. The LSD decreases and the performance increases in the frequency band of 0-4 kHz as the kernel size increases. However, in the frequency band of 4-8 kHz, the LSD of the restored waveform increases when the kernel size increases. Therefore, in the noise suppression process, the speech data was resampled at 8 kHz, and segmented into the length of 2048 samples each frame. The convolution kernel is set to 9×1, and the dilated factor was set to $2^{l-1}, l = [1, 2, ..., 8]$. The high-frequency reconstruction DNN is consisted of two LSTM layers and two fully-connected layers (see Fig. 6), and there are 1024 units for both LSTM and fully-connected layers. Backpropagation through time (BPTT) [15] was employed to improve the efficiency of the model training, and the truncated length was set to 480. For both the noise suppression DNN and high-frequency reconstruction DNN, the optimizer Adam [13] was applied.

TABLE I
EXPERIMENTAL CONDITIONS.

Environment	Sound-proof room
Ambient noise level	20.8 dB
Sampling frequency	16 kHz
Quantization bit rate	16 bits
Data	TIMIT Acoustic Phonetic Continuous Speech Corpus 9,000 files (4 hours) for training 240 files (12 minutes) for evaluation
Equipment	Polytec NLV-2500-5
Vibrating object	Pet-bottle

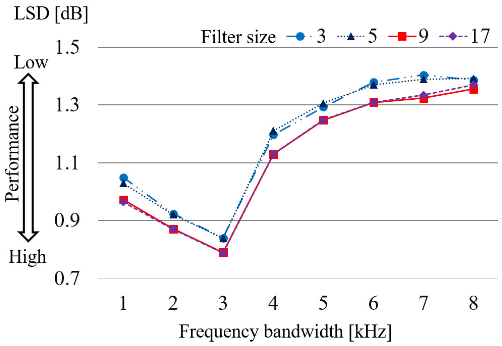


Fig. 9. Result of enhanced speech at each frequency bandwidth with different size convolution kernel.

TABLE II
OBJECTIVE EVALUATION RESULTS OF EACH METHOD.

	PESQ score	LSD [dB]	STOI score
(b)	1.76±0.40	1.62±0.10	0.85±0.04
(c)	2.18±0.49	1.09±0.09	0.90±0.02
(d)	2.00±0.30	1.21±0.09	0.93±0.03
(e)	2.35±0.50	1.11±0.08	0.94±0.03

(b),(c),(d),(e) are the observed speech, and the results of conventional STFT-based DNN, conventional waveform-based DNN, proposed waveform-based DNN

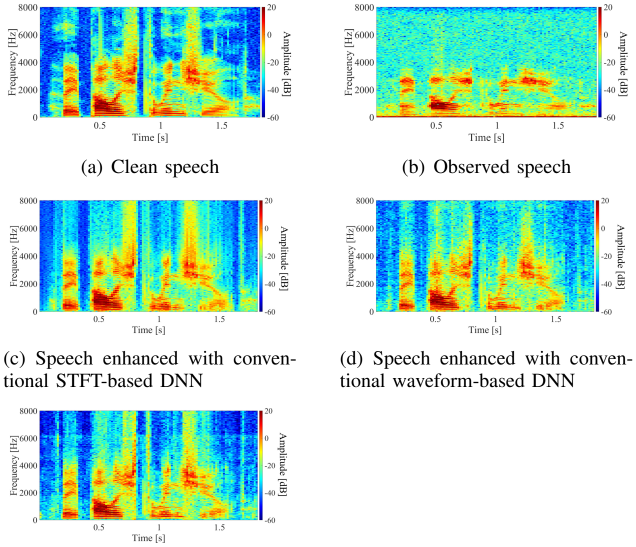
C. Experimental results

In order to prove the validity of the proposed method, the proposed method is compared with two conventional methods. One is the STFT-based DNN. It learned the relationship of the power spectrum between the observed speech and the clean speech, and used the phase of the observed speech in the inverse Fourier transform. The other is the waveform-based DNN, which used the same network as the noise suppression process of the proposed method, except that it has not been processed by high-frequency reconstruction. The observed speech and the results of each method were compared by objective evaluation experiments. The wideband perceptual evaluation of speech quality (PESQ), LSD, and short-time objective intelligibility (STOI) were adopted as objective measurements. The spectrograms for each method are shown in Fig. 10, and the results of the evaluation experiment are shown in Table II.

Figure 10 shows that the proposed method performs better in both noise suppression and high-frequency reconstruction than the conventional methods. From Table II, for the PSEQ and STOI scores, the proposed method achieved the best performance on both evaluation standards. For LSD, the proposed method is not as good as the conventional spectrum-based DNN but it is still about 0.51 dB higher than the observed speech. It can be concluded from Table II that the spectrum-based DNN can well map the power spectrum relationship between the observed signal and the target signal, but the quality of the generated speech is degraded due to the artifacts in the phase spectrum. Also, the DNN based on waveforms cannot extract features well due to the various kinds of distortion. The proposed method guarantees the phase information of the speech, and it reduces the noise of low-frequency components to improve the accuracy of the generated high-frequency information.

VI. Conclusion

In this paper, we proposed a speech enhancement method based on waveforms for the optical laser microphone. Since the optical laser microphone uses laser light to measure vibrations caused by sound waves, various kinds of distortion are mixed due to the vibration characteristics of the measurement object. However the conventional methods cannot adequately reduce these distortions. In this paper, we proposed a method based on a DNN with noise reduction and high-frequency band reconstruction processing. Through objective evaluation experiments, we confirmed that the proposed method outperforms the conventional method in PESQ, LSD, and STOI.



(a) Clean speech

(b) Observed speech

(c) Speech enhanced with conventional STFT-based DNN

(d) Speech enhanced with conventional waveform-based DNN

(e) Speech enhanced with proposed waveform-based DNN

Fig. 10. Enhanced speech with each method

In the future, we will compare the effects of different loss functions on the accuracy of the high-frequency reconstruction to determine a suitable loss function. Also, we will study speech enhancement for the optical laser microphone using different measurement objects and attempt to increase the distance between the LDV and object.

Acknowledgment

Part of this research was supported by JSPS Grants-in-Aid for Scientific Research (JP18K19829, JP19H04142) and Ritsumeikan Global Innovation Research Organization (R-GIRO).

References

[1] M.A. Clark, "An acoustic lens as a directional microphone," *Trans. IRE Prof. Group Audio,* vol. 25, no. 6, pp. 1152-1153, Jan. 1953.

[2] H. Ze, A.Senior, and M.Schuster, "Statistical parametric speech synthesis using deep neural networks," *Proc. IEEE ICASSP,* pp. 7962-7966, Oct. 2013.

[3] L.H. Chen, Z.H. Ling, L.J. Liu, and L.R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. Audio, Speech, Lang. Process.,* vol. 22, no. 12, pp. 1859-1872, Dec. 2014.

[4] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets." *Proc. Interspeech,* pp. 369-372, Aug. 2013.

[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *Proc. Interspeech,* 2013, pp. 436-440, Aug. 2013.

[6] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.,* vol. 23, no. 1, pp. 7-19, Jan. 2015.

[7] Y. Gu, Z.H. Ling, and L.R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," *Proc. Interspeech,* pp. 297-301, Sept. 2016.

[8] Z.H. Ling, Y. Ai, Y. Gu, and L.R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Trans. Audio, Speech, Lang. Process.,* vol. 26, no. 5, pp. 883-894, May. 2018.

[9] A.V.D. Oord et al., "WaveNet: A generative model for raw audio," arXiv:1609.03499, 2016.

[10] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *Proc. IEEE ICASSP,* pp. 5069-5073, Apr. 2018.

[11] ITU-T, "Recommendation G. 711: Pulse code modulation (PCM) of voice frequencies," *Int. Telecommun.* Union, 1988.

[12] C. Ledig, et al., "Photo-realistic single image super-resolution using a generative adversarial network," *Proc. IEEE CVPR,* pp. 105-114, Jul. 2017.

[13] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int conf. Learning Representation,* May 2015.

[14] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N. L. Dahlgren, "Acoustic-phonetic continuous speech corpus CD-ROM NIST speech disc 1-1.1," *NASA STI/Recon Tech.* Rep. LDC93S1, vol. 93, 1993.

[15] P.J. Werbos, "Backpropagation through time: what it does and how to do it," *Proc. IEEE,* vol. 78, no. 10, pp. 1550-1560, Oct. 1990.