PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Sound quality improvement for speech acquisition based on deep learning and harmonic reconstruction with laser microphone

Shoji UEDA[1]; Kenta IWAI[2]; Takahiro FUKUMORI[3]; Takanobu NISHIURA[4]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

[2, 3, 4] College of Information Science and Engineering, Ritsumeikan University, Japan

ABSTRACT

A laser microphone has recently been focused on acquiring the distant target speech without unnecessary sounds. It measures the vibration of the object near the target sound source by irradiating the object surface with the laser beam. However, the speech acquired with this microphone degrades the sound quality. For instance, the speech components at higher frequencies are attenuated by vibration characteristic of the object, and the speech is collapsed by the stationary noise due to lower intensity of the laser beam reflected from the object. To improve the sound quality of the degraded speech, deep neural networks (DNNs) have recently been proposed. They are trained by using a set of acoustic features extracted from the degraded speech and the clean speech. However, the speech components at higher frequencies are still attenuated after processed by the DNN. Therefore, we propose the method to improve the sound quality with harmonic reconstruction after processed by the DNN. The proposed method is based on the harmonic structure of the speech signal and reconstructs the speech harmonics by utilizing a non-linear function. We evaluated the effectiveness of the proposed method through perceptual evaluation of speech quality.

Keywords: Laser microphone, Deep learning, Deep neural network, Harmonic reconstruction, Sound quality improvement

1. INTRODUCTION

When the distant target speech is acquired with the air-conduction microphone, unnecessary sounds are also acquired because the diaphragm of the microphone is vibrated by not only the target speech but also unnecessary sounds. Recently, a laser microphone has been focused on the solution of this problem [1]. It measures the object's vibration by irradiating the object surface with the laser beam. Because of the measurement using the laser beam, it can measure the only target speech as far as the laser beam reaches the object. However, the speech acquired with the laser microphone degrades the sound quality. For example, the speech components at higher frequencies are attenuated, and the speech is collapsed by the stationary noise. These degradations are caused by vibration characteristic of the object and lower intensity of the laser beam reflected from the object. By these degradations, the speech harmonics at higher frequencies are lost, and the acquired speech becomes unclear. Therefore, it is necessary to improve the sound quality of the speech acquired with the laser microphone.

Recently, deep learning has attracted attention as the sound quality improvement of the degraded speech [2-4]. The basic strategy of applying a deep neural network (DNN) for the sound quality improvement is to train a network using a large set of acoustic features extracted from the degraded speech and the clean speech. The trained DNN can reduce the nonstationary noise as well as the stationary noise. However, if the input signal is close to zero, the output signal is also close to zero. Hence, when the DNN is trained using the speech with high-frequency attenuation like the speech acquired with the laser microphone, the DNN can ineffectively reconstruct the speech components at higher frequencies. Moreover, the larger the size of the DNN is, the larger the computational complexity of applying the DNN is.

In this paper, we propose a method for sound quality improvement to reconstruct the harmonic

[1] is0247es@ed.ritsumei.ac.jp
[2] iwai8sp@fc.ritsumei.ac.jp
[3] fukumori@fc.ritsumei.ac.jp
[4] nishiura@is.ritsumei.ac.jp

structure of the speech after processed by the DNN. The proposed method is based on the harmonic structure of the speech signal and reconstructs the speech harmonics by utilizing a non-linear function. To reconstruct the harmonic structure, we focus on that the periodic square wave which composes the harmonics of the fundamental frequency [5, 6]. The convolution between the spectrum of the acquired speech and that of the periodic square wave can reconstruct the speech harmonics. Then, the sound quality of the speech signal is improved, in other words, the speech signal is clearer. Moreover, the reconstruction of the harmonic structure can be employed with small configuration of the filter, and the proposed method has small computational complexity. Therefore, we can reduce the computational complexity for the sound quality improvement by combining the smaller DNN with the harmonic reconstruction.

2. SPEECH ACQUISITION WITH LASER MICROPHONE

In this section, we describe the degradation of the speech acquired with the laser microphone. Here, we define the laser microphone as the system for acquiring the sound by irradiating the surface of the object near the sound source with the laser beam. In this paper, we use a laser Doppler vibrometer (LDV) as the laser microphone. It can measure the vibration of the object which is vibrated slightly by the sound. It can be used to acquire the distant speech by irradiating the distant object with the laser beam. However, the speech acquired with the laser microphone is degraded by various noises such as the stationary noise [7] and the speckle noise [1, 8]. The stationary noise occurs when the intensity of the laser beam reflected from the object becomes lower. The spectral amplitude of this noise is proportional to the frequency. The speckle noise occurs in the vibration measurements on rough surfaces which are moved to vertical direction with respect to the irradiating direction. It has spikes which direct towards zero in the time domain. Moreover, the speech components at higher frequencies of the acquired speech are attenuated. This is caused by vibration characteristic of the object that the laser microphone irradiates with the laser beam.

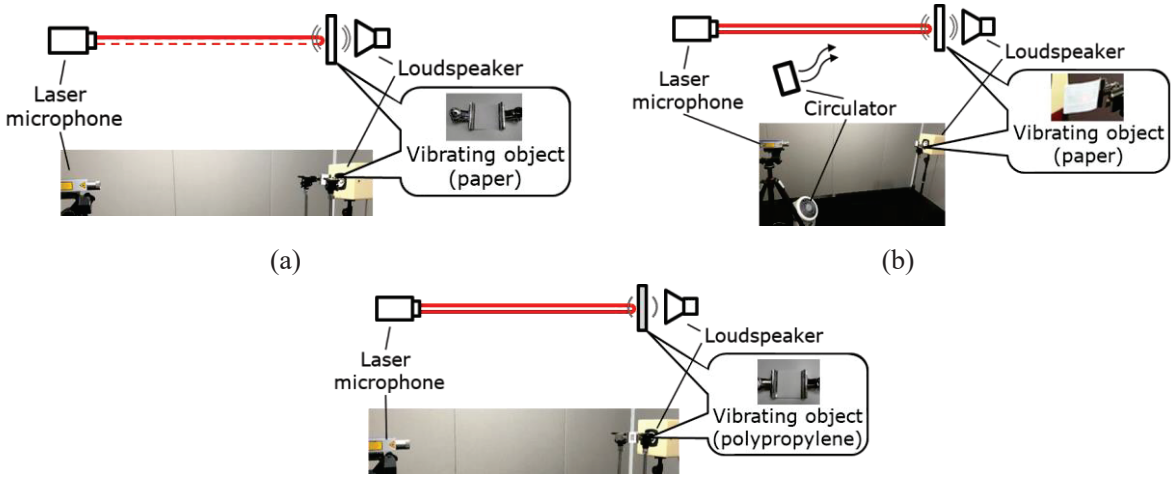


Figure 1 – Arrangement for measurement with the laser microphone using (a) a fixed paper, (b) a paper moved by the wind and (c) a polypropylene plate with a reflective tape

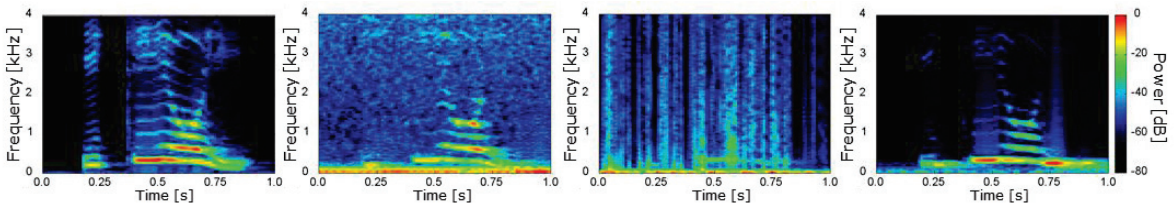


Figure 2 – Spectrograms of (a) clean speech, acquired speech by the laser microphone using (b) a fixed paper, (c) a paper moved by the wind and (d) a polypropylene plate with a reflective tape

Here, we show the measurement results to confirm the degradation of the speech acquired with the laser microphone. In this experiment, we focused on the degradation by the stationary noise, the speckle noise and high-frequency attenuation. Experimental setup is shown in Fig. 1. To confirm the stationary noise, we used a paper located in front of the loudspeaker. A paper reflects the laser beam with lower intensity. To confirm the speckle noise, we used a paper moved by the wind from the circulator. A paper which has rough surfaces is moved by the wind to vertical direction with respect to the irradiating direction. To confirm high-frequency attenuation, we used a polypropylene plate with a reflective tape. The vibration measured by a polypropylene plate has lower power at higher frequencies [9]. A reflective tape is used for suppressing the noise in the measurement with the laser microphone. We acquired Japanese speech "ikioi (Japanese pronunciation: ikioi)" that the female uttered in ATR phoneme balanced 216 words [10]. The spectrograms of acquired speech signals are shown in Fig. 2. From Fig. 2, we can see that the speeches acquired with the laser microphone are degraded. The acquired speech with the fixed paper is collapsed by the stationary noise over the speech components at higher frequencies. The speech acquired by using the paper moved by the wind is collapsed by the speckle noise which randomly occurs on the entire frequency range. In the acquired speech by using the polypropylene plate with the reflective tape, the noise is reduced compared with the acquired speeches of the object without the reflective tape. On the other hand, the speech components at higher frequencies attenuates because of the vibration characteristic of the object. From these results, we confirmed the degradation by acquiring with the laser microphone. Thus, it is required to reduce the noises, whereas it is also required to reconstruct the speech components at higher frequencies of speech for increase of sound quality.

3. SOUND QUALITY IMPROVEMENT BASED ON DEEP NEURAL NETWORK

Recently, DNNs have become popular for speech quality improvement. They are trained by a set of acoustic features extracted from the degraded speech and the clean speech. Figure 3 shows the training of the regression based DNN framework. In this paper, our purpose is sound quality improvement for the speech acquired with the laser microphone. To achieve our purpose, the DNN should be trained to learn the mapping from the acquired speech features to the clean speech ones. To improve the sound quality, various speech features have been proposed [2-4]. In the speech quality improvement, they can be converted back to the waveform in the time domain. For instance, log-power spectrum is popular as an acoustic feature for the deep learning. It can effectively learn the speech components which have small power at higher frequencies because it is used on the logarithmic scale. To obtain the improved speech, the amplitude spectrum is combined with the phase spectrum of the acquired speech. In addition, short-time waveform is also used as acoustic feature for the deep learning. It can learn the mapping function including phase information. Because of this, the DNN trained using short-time waveform can directly obtain the improved speech without phase replacement. For effective training the DNN, the concatenated frames are used as input to learn over a period of time.

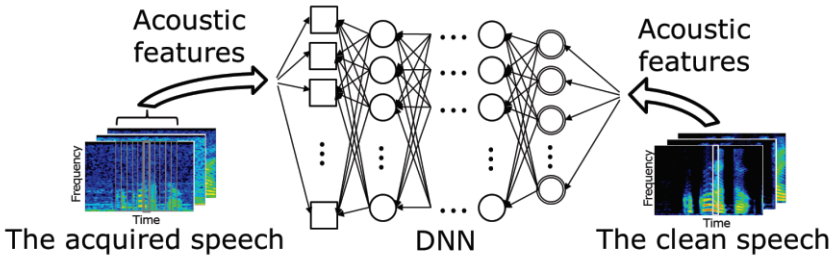


Figure 3 – Training of the regression based DNN framework

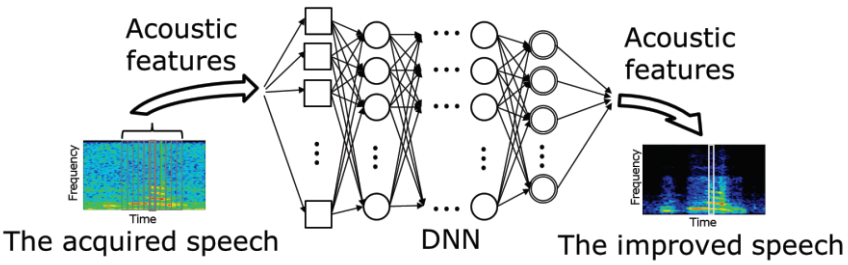


Figure 4 – Application of the regression based DNN framework

Figure 4 shows the application of the trained DNN. The clean speech features are estimated frame by frame from the acquired speech. However, the speech components at higher frequencies are still attenuated after processed by the DNN. This is because the output signal is close to zero when the input signal is close to zero. Because of this attenuation, the speech improved by the DNN becomes still unclear. To make the improved speech clear, the speech is required to reconstruct the components at higher frequencies. On the other hand, the larger the size of the DNN is, the larger the computational complexity of applying the DNN is. To suppress the computation load, it is necessary to utilize the function with simple configuration.

4. SOUND QUALITY IMPROVEMENT BASED ON DEEP NEURAL NETWORK AND HARMONIC RECONSTRUCTION

After the regression based DNN presented in the previous section, the noises acquired with the laser microphone are suppressed effectively. However, the processed speech still attenuates components at higher frequencies. This indicates that the speech loses harmonic structure which represents the speech characteristics. To solve this problem, we focus on that the harmonic structure composes the higher frequencies of the speech [11]. To suppress the computation load, we utilize the harmonic reconstruction with simple configuration of the filter. The noise reduction methods [5, 6] uses a non-linear function for harmonic reconstruction. By utilizing a non-linear function, the proposed method reconstructs the speech harmonics which are lost in the acquirement with the laser microphone.

4.1 Harmonic Reconstruction

The method to reconstruct the speech harmonics consists of applying a non-linear function NL[·] in the time domain. For example, the absolute value, minimum value, or maximum value function are used as the non-linear function. In this paper, we use a max function as the non-linear function NL[·] and the harmonic-reconstructed speech $x_{\mathrm{harmo}}(n)$ is calculated by

$$x_{\mathrm{harmo}}(n) = \mathrm{NL}[x(n)] \tag{1}$$

$$\mathrm{NL}[x(n)] = \max\{x(n), 0\} = x(n)p\big(x(n)\big) \tag{2}$$

$$p\big(x(n)\big) = \begin{cases} 1 & (\text{if } x(n) > 0) \\ 0 & (\text{if } x(n) < 0) \end{cases} \tag{3}$$

where $x(n)$ is the time waveform of the speech, and n is the discrete time index. Equation (2) is important to reconstruct the speech harmonics. Figure 5 shows an example of the harmonic reconstruction by Eq. (2). In Fig. 5, the acquired speech is extracted from the speech shown in Fig. 2 (d) with a center segment of 20 ms. In Fig. 5, the dotted lines are log-power spectra of the clean speech. From Fig. 5, we can see that the period of $p\big(x(n)\big)$ is same as the period of $x(n)$. Here, it is known that the periodic rectangular wave $q(n)$ is as following:

$$\begin{aligned} q(n) &= \sin(2\pi fn) + \frac{1}{3}\sin(6\pi fn) + \frac{1}{5}\sin(10\pi fn) + \cdots \\ &= \sum_{m=1}^{\infty} \frac{1}{2m-1}\sin\big((2m-1)\times 2\pi fn\big) \end{aligned} \tag{4}$$

where f is the normalized frequency of the rectangular wave. Hence, discrete-time Fourier transform (DTFT) of $q(n)$ is as following:

$$\begin{aligned} \mathrm{DTFT}[q(n)] &= \sum_{-\infty}^{\infty} q(n)e^{-j2\pi fn} \\ &= \frac{1}{N}\sum_{m=-\infty}^{\infty} R\big((2m-1)f\big)\delta(f-(2m-1)f) \end{aligned} \tag{5}$$

$$R\big((2m-1)f\big) = \mathrm{DTFT}\big[\sin\big((2m-1)\times 2\pi fn\big)\big] \tag{6}$$

where $\delta(m)$ is the Dirac distribution, $R(mf)$ is the DTFT of the elementary waveform at discrete frequency mf and N is the period of the rectangular wave. Equation (5) indicates that the

spectrum of the periodic square wave has the power in the components at the harmonics of discrete frequency $(2m-1)f$. Because of this characteristic, $p(x(n))$ has the harmonics which include the components at the higher frequencies. From Eqs. (1) and (2), the DTFT of $x_{\mathrm{harmo}}(n)$ is shown as

$$\mathrm{DTFT}[x_{\mathrm{harmo}}(n)] = \mathrm{DTFT}[x(n)] * \mathrm{DTFT}[p(x(n))] \tag{7}$$

Thus, the spectrum of the harmonic-reconstructed speech is the convolution between the spectrum of $x(n)$ and the spectrum of $p(x(n))$ which is the components of harmonics. $p(x(n))$ has the same fundamental frequency as $x(n)$. This is the reason why the harmonics are reconstructed at the accurate position with the convolution by Eq. (7). As shown in Fig. 5, the log-power envelope of $\mathrm{DTFT}[p(x(n))]$ is lower in higher frequencies. For the speech with high-frequency attenuation, the speech harmonics at higher frequencies are reconstructed using a few harmonics at lower frequencies. Because of this, the harmonic reconstruction using only Eq. (2) is ineffective to reconstruct the speech harmonics at the higher frequencies.

To solve this problem, we apply the pre-emphasis filter [12] in the proposed method. It enhances the components at higher frequencies to distribute the power evenly in the frequency domain. It is applied to $x_{\mathrm{harmo}}(n)$ to enhance the components at higher frequencies. It is calculated by

$$x'_{\mathrm{harmo}}(n) = x_{\mathrm{harmo}}(n) - \alpha x_{\mathrm{harmo}}(n-1) \tag{8}$$

where $x'_{\mathrm{harmo}}(n)$ is the enhanced speech and α is the filter coefficient. The filter coefficient α is generally set to 0.97.

4.2 Harmonic Reconstruction after Process of the DNN

Figure 6 shows overview of the proposed method. First, the sound quality of the acquired speech is improved by processing the DNN which is trained using a set of acoustic features extracted from other acquired speech and the clean speech. Second, harmonic reconstruction is applied to the speech processed by the DNN. To avoid over-enhancement of the reconstructed harmonics at higher frequencies, we add the processed speech to the harmonic-reconstructed speech. This process can enhance the components at frequencies which cannot be enhance by the pre-emphasis filter. In the process of harmonic reconstruction, the speech harmonics are reconstructed based on Section 4.1. As shown in Eqs. (2) and (8), harmonic reconstruction in the proposed method can suppress the computation load.

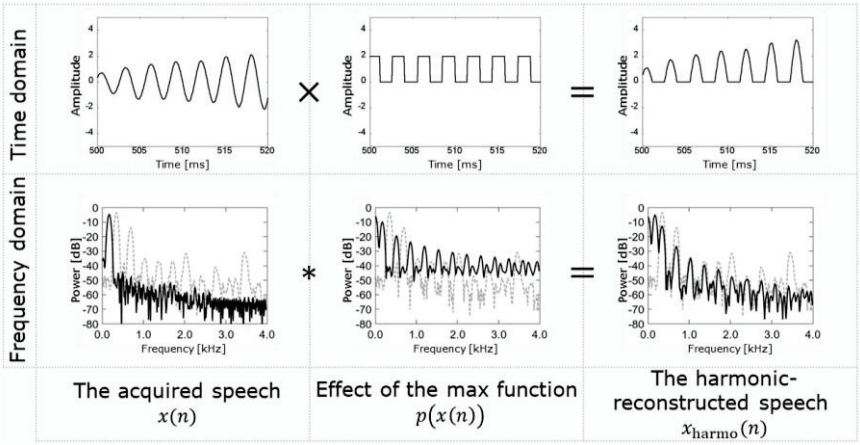


Figure 5 – The time waveform and log-power spectrum of the acquired speech $x(n)$, effect of the non-linear function $p(x(n))$ and harmonic-reconstructed speech $x_{\mathrm{harmo}}(n)$ (solid line) and log-power spectrum of the clean speech (dotted line)

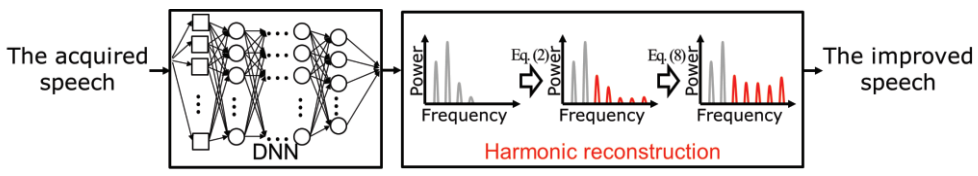


Figure 6 – Overview of the proposed method

5. EVALUATION EXPERIMENT ABOUT THE SOUND QUALITY OF THE IMPROVED SPEECH

5.1 Experimental Setup

We carried out an evaluation experiment to verify the effectiveness of the proposed method. In this experiment, we compared the sound qualities of the acquired speeches without any methods, with harmonic reconstruction, with the DNN, and with the proposed method. In this experiment, we carried out the measurement with the laser microphone to prepare the training and test data. Figure 7 and Table 1 respectively shows the arrangement and the measurement conditions. We used 7-hour utterances (5,024 utterances = 50 speakers × 2 genders × 50 or 53 sentences) from ATR phoneme balanced sentences [13] for training data and 1-hour utterances (3,888 utterances = 9 speakers × 2 genders × 216 words) from ATR phoneme balanced words [10] for test data.

In this experiment, the DNN used a feed-forward architecture. To verify the effectiveness of the number of the hidden layers, the DNN was used with 2 layers and 10 layers. We extracted log-power spectrum and short-time waveform from the speech with 256-points frame. For the input units of the DNN, they were used concatenated frames along with left and right of 11 frames. For the input and output units, the dimension of log-power spectrum was 257 and the dimension of short-time waveform was 256. Log-power spectrum was calculated by 512-point discrete Fourier transform with zero padding. The activation functions for the hidden units and for the output units were Rectified Linear Units (ReLUs) [14] and the sigmoid functions, respectively. The number of the hidden units in each layer was 2,048.

To evaluate the sound quality of the improved speech, we used perceptual evaluation of speech quality (PESQ) [15]. PESQ is the quality measure that correlates with the subjective evaluation and shows that the larger its value is, the higher the sound quality of speech is. The range of its value is -0.5 ~4.5.

5.2 Experimental Results

Figure 8 shows the spectrograms of the speech acquired with the laser microphone and the acquired speech processed by the harmonic reconstruction. As shown in Fig. 8 (a), the speech acquired with the laser microphone includes the stationary noise and the speckle noise. As shown in Fig. 8 (b), harmonics of the processed speech are ineffectively reconstructed because of the noises. Figure 9 shows the spectrograms of the speeches processed to the acquired speech by the DNN with 2 or 10 hidden layers. The speeches shown in Fig. 9 (a), (b) and Fig. 9 (c), (d) are trained using log-power spectrum and short-time waveform, respectively. From Fig. 9, we can see that the noises are reduced by the DNN trained by each training condition. However, the noise remains in the speech processed using short-time waveform. Figure 10 shows the spectrograms of speeches applied harmonic reconstruction to the speeches in Fig. 9. As shown in Fig. 10, the speech harmonics are effectively reconstructed in the red squares where the noise is reduced at higher frequencies.

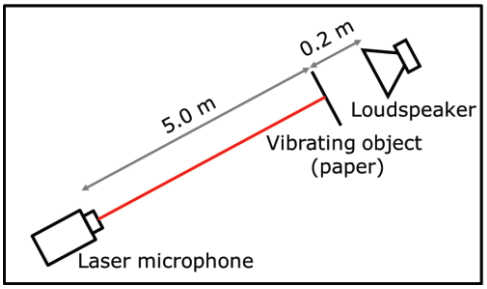
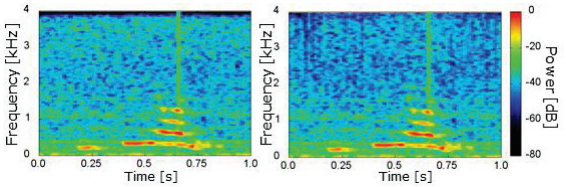


Figure 7 – Arrangement for training and test data in speech measurement with the laser microphone

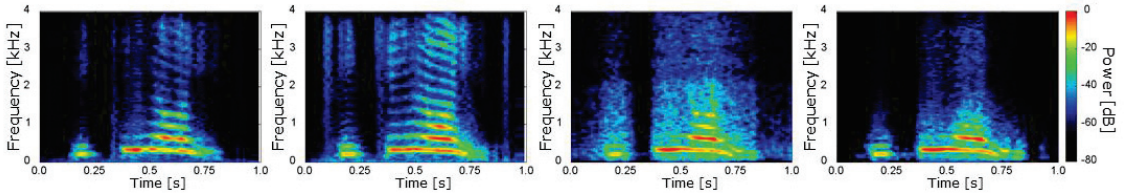
Table 1 – Experimental conditions for training and test data in speech measurement with the laser microphone

Environment	Conference room
Ambient noise level	33.4 dB
Output signal level	90 dB
Sampling frequency	8 kHz
Training data	7-hours utterances (5,024 utterances) in ATR phoneme balanced sentences
Test data	1-hour utterances (3,888 utterances) in ATR phoneme balanced words



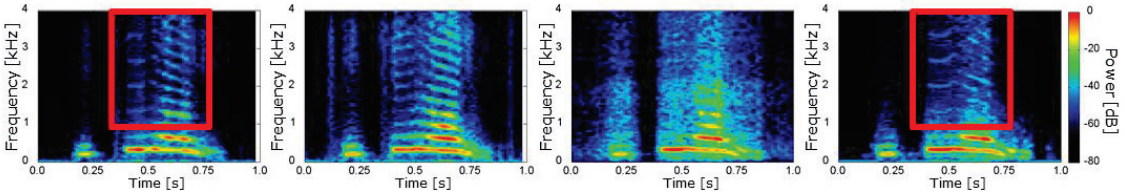
(a) (b)

Figure 8 – Spectrograms of (a) the speech acquired with the laser microphone and (b) the acquired speech processed by the harmonic reconstruction



(a) (b) (c) (d)

Figure 9 – Spectrograms of speeches processed by the trained DNN with (a) 2 and (b) 10 hidden layers using log-power spectrum, and (c) 2 and (d) 10 hidden layers using short-time waveform



(a) (b) (c) (d)

Figure 10 – Spectrograms of speeches applied harmonic reconstruction to the speeches in Figure 9

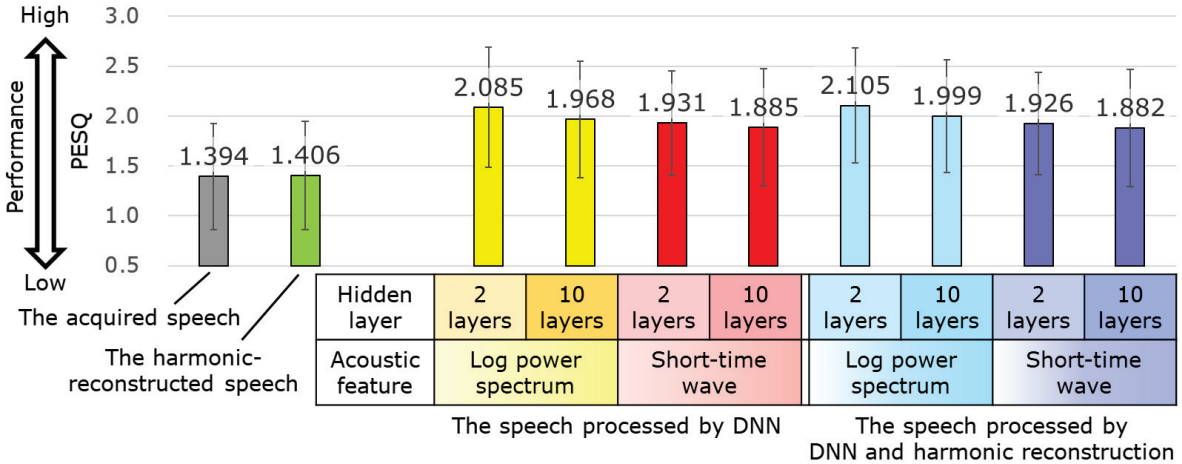


Figure 11 – PESQ scores

Figure 11 shows the evaluation result of PESQ. In Fig. 11, the horizontal axis shows the methods of speech quality improvement by using no processing, the deep learning and the proposed method. Also, the vertical axis shows the value of PESQ calculated from test data. Each bar shows the average values and the error bar shows the standard deviation. As shown in Fig. 11, the PESQ of the proposed method is larger than that of the deep learning when the DNN is trained using the log-power spectrum. On the other hand, the PESQ of the proposed method is smaller than that of the deep learning when the DNN is trained using the short-time waveform. This is because the proposed method can ineffectively reconstruct the speech harmonics by the noise around the speech harmonics at low frequencies. To solve this problem, we need to reconstruct the speech harmonics from the speech with this noise. Furthermore, we need to prepare the more training data for the DNN which effectively performs to reduce the noises.

6. CONCLUSIONS

In this paper, we propose the method of the speech quality improvement with deep learning and harmonic reconstruction for the speech acquired with the laser microphone. In the proposed method, the non-linear function is applied to reconstruct the speech harmonics and the pre-emphasis filter is applied to enhance the low power at higher frequencies. We evaluated the sound quality of the improved speech by PESQ while changing acoustic features and the number of the hidden layers in the DNN. As the result, the proposed method using log-power spectrum is more effective than the proposed method using short-time waveform. For improvement of the DNN using short-time waveform, it is necessary to reduce the noise around the speech harmonics at low frequencies. In the future, we intend to modify the harmonic reconstruction such as adjusting the power of the harmonic reconstruction by analyzing the tendency of the attenuation and the noise. Furthermore, we will improve the performance of the DNN by preparing the larger training set.

ACKNOWLEDGEMENTS

This work was partly supported by JST COI, JSPS KAKENHI Grant Numbers JP18K19829 and JP19H04142.

REFERENCES

1. Y. Avargel and I. Cohen, "Speech measurements using a laser Doppler vibrometer sensor: Application to speech enhancement," in *Proc. 3rd HSCMA*, Edinburgh, UK, Jun. 2011, pp. 109-114.
2. A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, pp. 7092-7096, 2013.
3. X. lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. 14th INTERSPEECH*, pp. 3444-3448, 2013.
4. R. K. Srivastava, K. Greff and J. Schmidhuber, "Training very deep networks," in *Proc. 28th NIPS*, pp. 2377-2385, 2015.
5. C. Plapous, C. Marro and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098-2108, 2006.
6. H. Xuchu and Z. Xiaojing, "Speech enhancement using harmonic regeneration," in *Proc. ICASSP*, pp. 150-152, 2011.
7. M. Johansmann, G. Siegmund and M. Pineda, "Targeting the limits of laser Doppler vibrometry," in *Proc. IDEMA*, pp. 1-12, 2005.
8. J. Vass, R. B. Randall, C. Cristalli, B. Torcianti, P. Sovka and R. Smid, "Spectrum enhancement of laser vibrometer signals using a new speckle noise reduction technique and angular resampling," in *Proc. 14th ICSV*, pp. 636-644, 2007.
9. G. W. Chantry, J. W. Fleming, G. W. F. Pardoe, W. Reddish and H. A. Willis, "Absorption spectra of polypropylene in the millimetre and submillimetre regions," *Infrared Physics*, vol. 11, no. 2, pp. 109-118, 1971.
10. H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa and T. Watanabe, "Construction of a large-scale Japanese speech database and its management system," in *Proc. ICASSP*, pp. 560-563, 1989.
11. H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.
12. Z. B. Nossair, P. L. Silsbee and S. A. Zahorian, "Signal modeling enhancements for automatic speech recognition," in *Proc. ICASSP*, pp. 824–827, 1995.
13. A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano. "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357-363, 1990.
14. V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th ICML*, pp. 807-814, 2010.
15. A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, pp. 749-752, 2001.