Deep-Learning-Based Speech Enhancement with Rough-Focused Optical Laser Microphone by Reconstructing Complex Spectrum

Yuki Nakano*, Yuting Geng*, Kenta Iwai*, Takanobu Nishiura*
* Ritsumeikan University, Osaka, Japan
E-mail: {is0570he@ed, geng@fc, iwai18sp@fc, nishiura@is}.ritsumei.ac.jp

***Abstract*—Rough-focused recording with an optical laser microphone allows for recording that is wide ranging and robust against changes in the position of a vibrating object. However, the recorded speech suffers from noise due to laser diffusion and missing signal components. To solve this problem, we propose a speech enhancement method for rough-focused recordings that reconstructs a complex spectrum. The conventional speech enhancement method for rough-focused optical laser microphones reconstructs only an amplitude spectrogram without considering phase components, resulting in lower quality speech enhancement. In contrast, the proposed method simultaneously reconstructs both amplitude and phase components of rough-focused recordings by reconstructing a complex spectrum. We compared the method with the conventional method. The results show that the proposed method performed equivalently to or better than the conventional method in objective evaluations.**

I. Introduction

The optical laser microphone [1] has attracted attention for use in acoustic systems capable of recording target speech from a distance [2], [3]. It measures speech-induced vibration by irradiating a laser beam onto the surface of a vibrating object and captures the speech. Generally, acoustic measurements with this microphone have been conducted with a focused laser beam to obtain sufficient reflected light. In contrast, a recording method that operates in an out-of-focus state, called rough-focused recording [4], has been studied. In rough-focused recording, the optical laser microphone irradiates the laser beam over a wide area, allowing for recording that is both wide-ranging and robust against changes in the position of a vibrating object. However, the wide irradiation area of the laser beam results in the reflection of the laser becoming diffused when measuring acoustic signals, causing noise and missing signal components in the recorded signal. Therefore, noise suppression and reconstruction of missing signal components are the main tasks of speech enhancement for the rough-focused optical laser microphone.

To address this issue, deep-learning-based speech enhancement for rough-focused recording has been proposed [4]. In [4], a single model enables speech enhancement for recorded speech in different rough-focus conditions. Then, the model reconstructs an amplitude spectrum of clean speech from one of recorded speech. Therefore, the speech enhancement performance in [4] is inadequate in that only the amplitude spectrum of the speech signal is reconstructed while the phase spectrum is not taken into account.

In this paper, we propose a deep-learning-based speech enhancement for the rough-focused optical laser microphone that reconstructs a complex spectrum and discuss the feasibility of speech enhancement for rough-focused recording with simultaneously reconstructed both amplitude and phase components. Since the complex spectrum contains both amplitude and phase components, the reconstruction of the spectrum allows for phase-aware speech enhancement. The complex-spectrum reconstruction model is designed on the basis of a previous study [5]. To improve the performance in reconstructing signal components in the high-frequency band, we also investigate the design of the loss function.

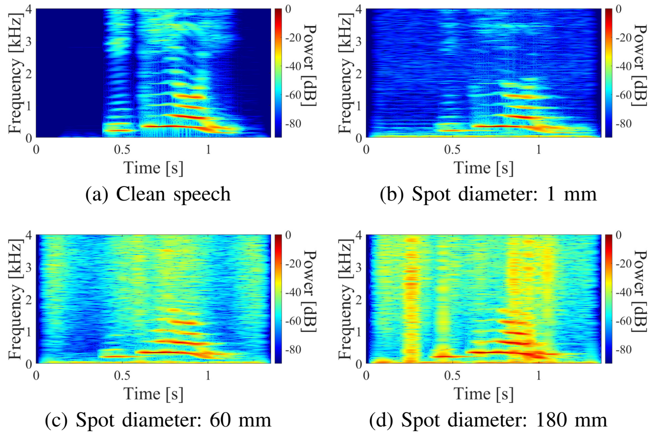
The remainder of the paper is organized as follows. In Section II, we describe speech-quality degradation with the rough-focused optical laser microphone and related work. In Section III, we present the proposed method. In Section IV we present the experiments and the results. Finally, we conclude the paper in Section V.

II. Speech-Quality Degradation in Rough-Focused Optical Laser Microphone

In a rough-focused optical laser microphone, recording robust against movements of vibrating objects comes at the cost of degradation in speech quality, as discussed below. In this paper, we use the spot diameter of the laser beam on a vibrating object as a metric to represent the state of rough focus.

Spectrograms of clean speech and recorded speech with the rough-focused optical laser microphone are shown in Fig. 1. Fig. 1 (b) shows that signal components in the high-frequency band (2 to 4 kHz) are missing. Also, Fig. 1 (d) shows that the recorded speech contains impact and stationary noises. This is attributed to the decrease in the intensity of reflected light due to rough-focus recording, resulting in measurement errors at the detector and subsequent noise contamination [6]. Then, Fig. 1 shows that the larger the spot diameter, the greater the degradation in sound quality.

Therefore, to obtain high-quality speech with the rough-



(a) Clean speech

(b) Spot diameter: 1 mm

(c) Spot diameter: 60 mm

(d) Spot diameter: 180 mm

Fig. 1: Spectrograms of recorded speech with rough-focused optical laser microphone.

focused optical laser microphone, it is necessary to suppress the noise and enhance the high-frequency components of the speech through speech enhancement.

To address the above problems, a previous study [4] proposed deep neural network (DNN)-based speech enhancement by reconstructing a log-power spectrum for rough-focus recorded speech. In the method [4], rough-focus recorded speech $\mathbf{x} \in \mathbb{R}^T$ is subjected to a short-time Fourier transform (STFT), and the complex spectrum $\mathbf{X} \in \mathbb{C}^{K \times L}$ is calculated as

$$\mathbf{X} = \mathrm{STFT}[\mathbf{x}], \tag{1}$$

where T is the number of samples, L is the number of frames, K is the number of frequency bins, and $\mathrm{STFT}[\cdot]$ represents the STFT operator, respectively. Then, the amplitude spectrograms and phase spectrograms of $\mathbf{X}$ are denoted as $|\mathbf{X}|$ and $\angle\mathbf{X}$. Next, the log-power spectrogram $10\log_{10}|\mathbf{X}|^2$ is input to DNN to obtain $10\log_{10}|\hat{\mathbf{Y}}|$, where $10\log_{10}|\mathbf{X}|^2$ is the log-power spectrogram of $\mathbf{X}$, and $10\log_{10}|\hat{\mathbf{Y}}|^2$ is the log-power spectrogram of enhanced speech. Finally, the amplitude spectrogram $|\hat{\mathbf{Y}}|$ is calculated from $10\log_{10}|\hat{\mathbf{Y}}|^2$, and enhanced speech $\hat{\mathbf{y}} \in \mathbb{R}^T$ is obtained by performing an inverse STFT (ISTFT) using $|\hat{\mathbf{Y}}|$ and $\angle\mathbf{X}$.

$$\hat{\mathbf{y}} = \mathrm{ISTFT}[|\hat{\mathbf{Y}}|e^{j\angle\mathbf{X}}], \tag{2}$$

where $\mathrm{ISTFT}[\cdot]$ represents the ISTFT operator. In (2), the degraded phase spectrum of the roughly recorded speech is used as is to calculate $\hat{\mathbf{y}}$. Therefore, the speech enhancement performance is insufficient as phase components are not considered.

III. Proposed Speech Enhancement for Rough-Focused Optical Laser Microphone

A. Overview of proposed speech enhancement

We focus on the deep complex convolution recurrent network (DCCRN) [5] for its network architecture, which is designed to handle the complex spectrum. Since the complex spectrum contains amplitude components and phase components, DCCRN enables phase-aware speech enhancement. An overview of the proposed method is shown in Fig. 2. First, $\mathbf{x}$ is first subjected to STFT, and $\mathbf{X}$ is calculated. Next, $\mathbf{X}$ is input to DCCRN, and the complex spectrum of enhanced speech $\hat{\mathbf{Y}}$ is obtained. In the stage of model training, the error E_{Prop} is calculated using the proposed loss function from the enhanced speech $\hat{\mathbf{y}}$ and the clean speech $\mathbf{y}$. A detailed definition of the loss function is given in Section III-C. In the proposed method, the speech enhancement model with DCCRN is trained by minimizing the error calculated from E_{Prop} for speech recorded with a rough-focused optical laser microphone.

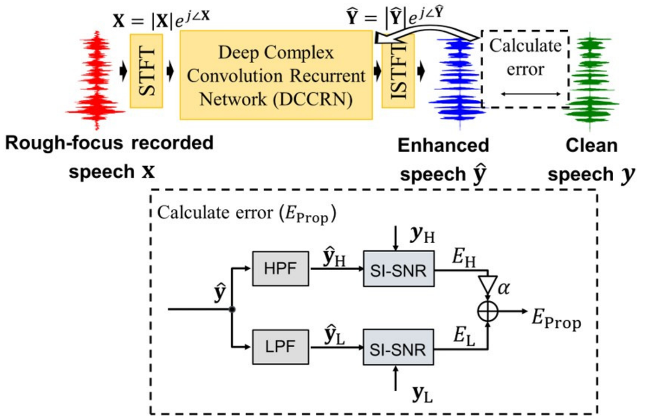

$\mathbf{X} = |\mathbf{X}|e^{j\angle\mathbf{X}}$
$\hat{\mathbf{Y}} = |\hat{\mathbf{Y}}|e^{j\angle\hat{\mathbf{Y}}}$
STFT
Deep Complex Convolution Recurrent Network (DCCRN)
ISTFT
Calculate error
Rough-focus recorded speech $\mathbf{x}$
Enhanced speech $\hat{\mathbf{y}}$
Clean speech $\mathbf{y}$
Calculate error (E_{Prop})
$\hat{\mathbf{y}}$
HPF
LPF
$\hat{\mathbf{y}}_{\mathrm{H}}$
$\hat{\mathbf{y}}_{\mathrm{L}}$
$\mathbf{y}_{\mathrm{H}}$
$\mathbf{y}_{\mathrm{L}}$
SI-SNR
SI-SNR
E_{H}
E_{L}
α
E_{Prop}

Fig. 2: Overview of training model in speech enhancement for recorded speech with rough-focused optical laser microphone.

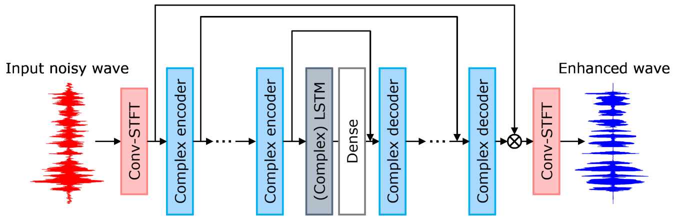


Fig. 3: DCCRN network.

B. Network architectures

The network architectures of DCCRN and the components of the complex encoder are shown in Figs. 3 and 4. DCCRN, originally described in [5], is an encoder-decoder architecture with two long short term memory (LSTM) layers. As shown in Fig. 4, the complex encoder consists of complex convolutional layers, complex batch normalization, and an activation function (PReLU). In the complex convolutional layers, complex-valued filters are convolved with the input on the basis of the rules of complex multiplication. The complex decoder is composed of the complex convolutional layers, complex batch normalization, and an activation function (PReLU) as in the complex encoder. Then, in the complex convolutional layers of the complex decoder, transposed convolution is performed.

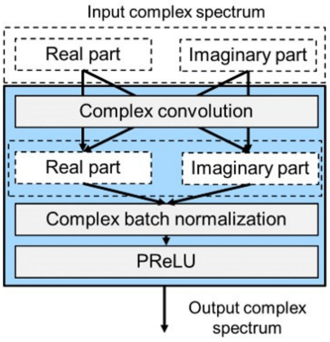


Fig. 4: Complex encoder.

C. Loss function

Generally, the loss function used in training of the DCCRN is the scale-invariant source-to-noise ratio(SI-SNR). SI-SNR is defined as

$$\begin{cases} \mathbf{y}_{\text{target}} & := \dfrac{\langle \hat{\mathbf{y}}, \mathbf{y} \rangle \cdot \mathbf{y}}{\|\mathbf{y}\|^2} \\ \mathbf{e}_{\text{noise}} & := \hat{\mathbf{y}} - \mathbf{y}_{\text{target}} \\ \text{SI-SNR} & := 10 \log_{10} \dfrac{\|\mathbf{y}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2}, \end{cases} \tag{3}$$

where $\mathbf{y}$ is clean speech, $\langle \cdot, \cdot \rangle$ denotes the inner product of vectors, and $\| \cdot \|$ denotes the Euclidean norm. Additionally, the loss $E_{\text{SI-SNR}}$ is expressed as

$$E_{\text{SI-SNR}} = -\,\text{SI-SNR}. \tag{4}$$

In this paper, we propose a method using loss function $E_{\text{SI-SNR}}$ and a method using the proposed loss function E_{Prop}. The method of calculating the proposed loss function E_{Prop} is explained as follows. First, $\hat{\mathbf{y}}$ is obtained from $\mathbf{x}$ through a series of processes based on the DCCRN. Next, high-pass and low-pass filters are applied to both $\hat{\mathbf{y}}$ and $\mathbf{y}$ in order to calculate $\hat{\mathbf{y}}_{\text{H}}$, $\hat{\mathbf{y}}_{\text{L}}$ and $\mathbf{y}_{\text{H}}$, $\mathbf{y}_{\text{L}}$. In this context, $\hat{\mathbf{y}}_{\text{H}}$, $\hat{\mathbf{y}}_{\text{L}}$ and $\mathbf{y}_{\text{H}}$, $\mathbf{y}_{\text{L}}$ represent the high-frequency and low-frequency components of $\hat{\mathbf{y}}$ and $\mathbf{y}$. Then, from $\hat{\mathbf{y}}_{\text{H}}$, $\mathbf{y}_{\text{H}}$ and $\hat{\mathbf{y}}_{\text{L}}$, $\mathbf{y}_{\text{L}}$, the loss in the high and low frequency components E_{H}, E_{L} are calculated. Finally, the total loss is calculated as

$$E_{\text{Prop}} = \alpha E_{\text{H}} + E_{\text{L}}, \tag{5}$$

where α is the weight of E_{H}. By adjusting α, it is possible to correct the influence of the low-frequency and high-frequency components on $E_{\text{SI-SNR}}$. Given the missing components in the high-frequency band in speech recorded with a rough-focused optical laser microphone, we believe that using E_{Prop} as the loss function can help the model better focus on reconstructing high-frequency components.

IV. Experiments and Results

We conducted experiments to evaluate the performance of the proposed method. In the experiments, we refer to the training method with the original loss function as Prop. A, and we refer to that with the proposed loss function as Prop. B.

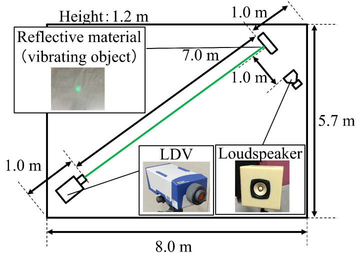


Fig. 5: Arrangement of measurement equipment.

TABLE I: Recording conditions.

Reverberation time T_{60}	300 ms
Ambient noise level L_{A}	29.7 dB
Sampling frequency	8 kHz
Quantization	16 bits
Temperature	23.8°C
Humidity	19.5%

TABLE II: Recording equipment.

Laser doppler vibrometer (LDV)	Polytec, VFX-F-110
A/D, D/A converter	RME, Fireface UFX
Loudspeaker	Fostex, FE83En
Loudspeaker amplifier	BOSE, 1705Ⅱ
Reflective material	ONOSOKKI, LV-0012

A. Experimental setups

We recorded a speech dataset using a rough-focused optical laser microphone. The dataset was divided into two parts to train and evaluate the speech enhancement model. Specifically, 503 Advanced Telecommunications Research (ATR) phoneme-balanced sentences [7] (average duration: approximately 11.8 seconds per file) were used as the training data, and 216 ATR phoneme-balanced words [8] (average duration: 2.1 seconds per file) were used as the evaluation data. The arrangement of the equipment used for recording the training and evaluation data, the recording conditions, and the recording equipment are shown in Fig. 5 and Tables I and II, respectively. In this recording, 15,368 utterances from datasets [7], which include 300 speakers (150 female and 150 male speakers), were recorded with an optical laser microphone in various focus settings (spot diameter: 1, 60, 120, and 180 mm) as the training data. Also, 3,024 utterances from datasets [8], which include 14 speakers (7 female and 7 male speakers) were recorded as the evaluation data. The experimental settings for training of DNN are shown in Table III.

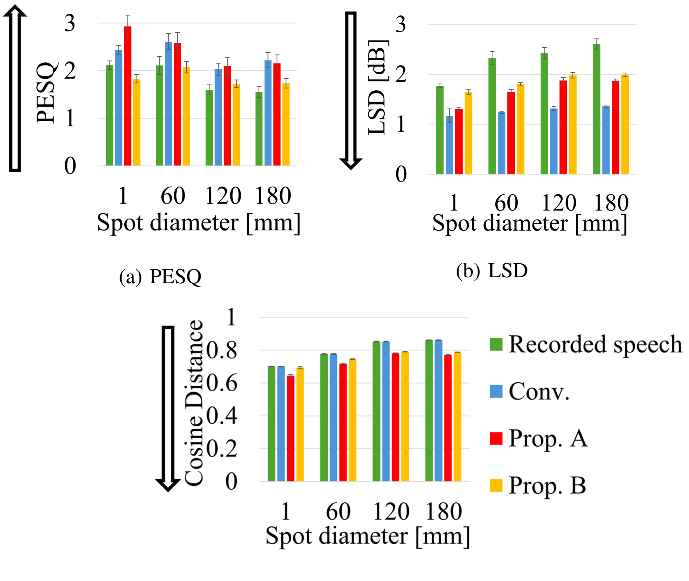
We used the perceptual evaluation of speech quality (PESQ) [9], log-spectral distance (LSD) [10], and cosine distance as the evaluation metrics. PESQ is used for speech in the range of 300–3400 Hz and is highly correlated with subjective evaluation of speech quality. The PESQ score ranges from -0.5 to 4.5, where higher scores indicate speech that is more intelligible to human auditory perception. LSD is a metric that shows magnitude distortion, where lower scores indicate higher quality of amplitude components in the evaluated speech. Cosine distance is a measure of the phase errors, defined as

$$\text{cosine distance} = 1 - \cos(\theta_{\text{Clean}} - \theta_{\text{Enh}}), \tag{6}$$

where θ_{Clean} is the phase spectrum of clean speech, and θ_{Enh} is the phase spectrum of enhanced and recorded speech. The

TABLE III: Experimental settings.

Sampling frequency	8 kHz
Quantization	16 bit
Window function	Hamming window
STFT length	512
Hidden layers	5 layers
LSTM layers	2 layers
Epochs	600
Batch size	32
Learning rate	0.0001
Optimizer	Adam
Loss function	$E_{\text{SI-SNR}}$ / E_{Prop}
Cutoff frequency in E_{Prop}	2 kHz
α in E_{Prop}	5

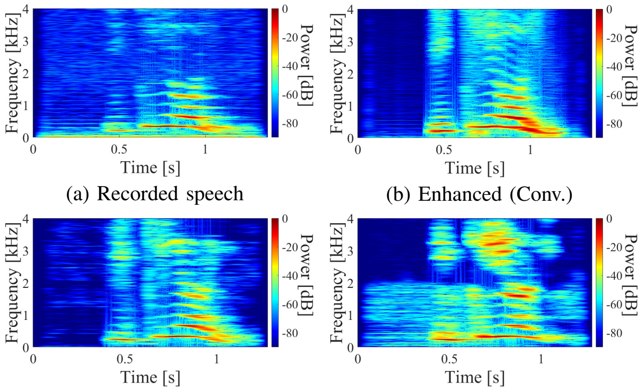


(a) PESQ

(b) LSD

(c) Cosine distance

Fig. 6: Evaluation results.



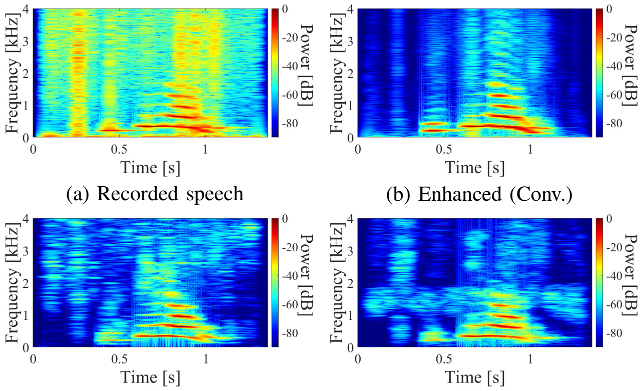
(a) Recorded speech

(b) Enhanced (Conv.)

(c) Enhanced (Prop. A)

(d) Enhanced (Prop. B)

Fig. 7: Spectrograms of recorded and enhanced speech (spot diameter: 1 mm)



(a) Recorded speech

(b) Enhanced (Conv.)

(c) Enhanced (Prop. A)

(d) Enhanced (Prop. B)

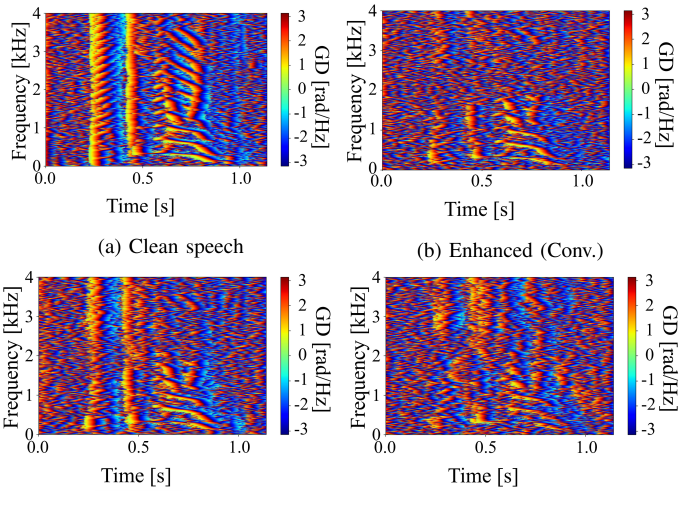
Fig. 8: Spectrograms of recorded and enhanced speech (spot diameter: 180 mm)

cosine distance score ranges from 0.0 to 2.0, where lower scores indicate higher quality of phase components in the evaluated speech.

B. Experimental results and discussion

The evaluation results in terms of PESQ, LSD, and cosine distance are shown in Fig. 6. From Fig. 6 (c), the cosine distance of recorded speech and the conventional method (Conv.) had the same score because enhanced speech with method [4] is calculated using STFT with the estimated amplitude and the phase of the recorded speech. Fig. 6 (a) shows that the PESQ score of the enhanced speech with Prop. A surpassed that of the enhanced speech with Conv. for the spot diameter of 1 mm. On the other hand, the PESQ score of the enhanced speech with Prop. B was lower than both the PESQ score of the enhanced speech with Prop. A and that with the conventional method. Therefore, the experimental results for PESQ show that Prop. A had the highest performance at a spot diameter of 1 mm, and Prop. A and Conv. performed equivalently at spot diameters other than 1 mm. Fig. 6 (b) shows that the LSD score of the enhanced speech with Prop. A was comparable to that with Conv. only for the spot diameter of 1 mm. The LSD score of the enhanced speech with Prop. B was higher than that with Conv. for all spot diameters. Therefore, Conv. and Prop. A had comparable amplitude reconstruction performance at the spot diameter of 1 mm. However, Conv. had the highest amplitude reconstruction performance at spot diameters other than 1 mm. Fig. 6 (c) shows that the cosine distance of the enhanced speech with Prop. A and that with Prop. B were lower than that with Conv. for all spot diameters. This indicates that Prop. A and Prop. B achieved phase reconstruction simultaneously with amplitude reconstruction at all spot diameters. Despite reconstructing both amplitude and phase simultaneously at all spot diameters, the PESQ score only improved with Prop. A at a spot diameter of 1 mm. This may be caused by the inferior amplitude reconstruction performance compared to Conv. at spot diameters other than 1 mm, as indicated in Fig. 6 (b) .

The spectrograms of the recorded speech and enhanced speech are shown in Figs. 7 and 8. Fig. 7 (b) and Fig. 8 (b)



(a) Clean speech

(b) Enhanced (Conv.)

(c) Enhanced (Prop. A)

(d) Enhanced (Prop. B)

Fig. 9: Group delay of clean speech and enhanced speech (spot diameter: 1 mm)

show that method [4] suppressed impact and stationary noises in the recorded speech and reconstructed the high-frequency speech components. This could be the reason for the good scores of method [4] in PESQ and LSD at all spot diameters. Fig. 7 (c) and Fig. 8 (c) show that Prop. A suppressed both impact and stationary noises and failed to reconstruct the speech components in the frequency band above 2 kHz. Given the low requirement for the reconstruction of the high-frequency components attributable to the slight missing components in speech recorded at a 1-mm-spot diameter, Prop. A was found to score well in PESQ and LSD at this diameter. Fig. 7 (d) and Fig. 8 (d) show that Prop. B suppressed noises in the frequency band above 2 kHz and reconstructed the high-frequency speech components. Given the higher performance in enhancing high-frequency components compared with Prop A, Prop. B shows the feasibility of improving the performance of reconstructing high-frequency components through E_{Prop}.

The group delay spectrograms [11] of clean and enhanced speech are shown in Fig. 9. It is known that the group delay and amplitude have strongly constrained relationships [12], and it can be confirmed that the group delay spectrogram has a similar structure to the log-power spectrogram in clean speech. Figs. 9 (a) and (b) show the distortion of phase components in the frequency band above 2 kHz. Fig. 9 (c) shows the reconstruction of the phase components in the frequency band above 2 kHz with Prop. A. It is clear that Prop. A achieves phase-aware speech enhancement. As shown above, the amplitude reconstruction of Conv. and Prop. A is accurate, while the phase reconstruction is done only by Prop. A. This could be the reason for the significant improvement in PESQ score of Prop. A.

In summary, the proposed method improves the quality of speech recorded with a rough-focused optical laser microphone. However, except for the case with a spot diameter of 1 mm, the performance of Prop. A is equivalent to or slightly lower than that of the conventional method. The reason for this is that the amplitude components could not be sufficiently reconstructed as a result of considering the phase components. Additionally, the enhanced speech with Prop. B, which modified the loss function to reconstruct high-frequency components, demonstrates the feasibility of improving the performance of reconstructing high-frequency components through E_{Prop}.

V. Conclusion

We proposed a DCCRN-based speech enhancement method for the rough-focused optical laser microphone. We evaluated our method by comparing PESQ, LSD, and cosine distance. Objective experiments showed that the proposed method, which considers phase components, is superior to the conventional method. In the future, we aim to achieve further high-quality speech for the rough-focused optical laser microphone by modifying the loss function and using features other than the complex spectrum.

VI. Acknowledgements

This work was partly supported by the Ritsumeikan Advanced Research Academy (RARA), Ritsumeikan Global Innovation Research Organization (R-GIRO), and by JSPS KAKENHI Grant Numbers JP21H03488, JP23H03425, JP23K21691, JP23K28115, and JP24K20803.

References

[1] J. A. Bucaro and N. Lagakos, "Lightweight fiber optic microphones and accelerometers," *Review of Scientific Instruments*, vol. 72, no. 6, pp. 2816–2821, 2001.

[2] T. Fukumori, C. Cai, Y. Zhang, L. E. Hafi, Y. Hagiwara, and T. Nishiura, "Optical laser microphone for human-robot interaction: Speech recognition in extremely noisy service environments," *Advanced Robotics*, vol. 36, no. 5-6, pp. 304–317, 2022.

[3] Y.-M. Lin, J.-Y. Han, C.-H. Lin, and Y.-H. Lai, "Optical microphone-based speech reconstruction system with deep learning for individuals with hearing loss," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 12, pp. 3330–3341, 2023.

[4] K. Miyazato, W. Haonan, K. Iwai, and T. Nishiura, "Study on speech quality improvement of rough pint recorded by optical laser microphone (in Japanese)," *2022 Spring Meeting of Acoustical Society of Japan*, pp. 403–406, 2022.

[5] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proceedings Interspeech 2020*, 2020, pp. 2472–2476.

[6] M. Johansmann, G. Siegmund, and M. Pineda, "Targeting the limits of laser Doppler vibrometry," *Proceedings IDEMA*, pp. 1–12, 2005.

[7] Y. Sagisaka, “A large-scale Japanese speech database,” *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, pp. 1089–1092, 1990.

[8] 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.

[9] ITU-T, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” *International Telecommunication Union Telecommunication Standardization Sector Recommendation*, no. P.862, 2001.

[10] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition.* (Prentice Hall signal processing series). Prentice Hall, 1993.

[11] N. B. Thien, Y. Wakabayashi, K. Iwai, and T. Nishiura, “Inter-frequency phase difference for phase reconstruction using deep neural networks and maximum likelihood,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1667–1680, 2023.

[12] F. Auger, É. Chassande-Mottin, and P. Flandrin, “On phase-magnitude relationships in the short-time Fourier transform,” *IEEE Signal Processing Letters*, vol. 19, no. 5, pp. 267–270, 2012.