PAPER

DNN-Based Low-Musical-Noise Single-Channel Speech Enhancement Based on Higher-Order-Moments Matching

Satoshi MIZOGUCHI[†a)], ***Nonmember***, **Yuki SAITO**[†b)], **Shinnosuke TAKAMICHI**[†c)], ***and*** **Hiroshi SARUWATARI**[†d)], ***Members***

SUMMARY We propose deep neural network (DNN)-based speech enhancement that reduces musical noise and achieves better auditory impressions. The musical noise is an artifact generated by nonlinear signal processing and negatively affects the auditory impressions. We aim to develop musical-noise-free speech enhancement methods that suppress the musical noise generation and produce perceptually-comfortable enhanced speech. DNN-based speech enhancement using a soft mask achieves high noise reduction but generates musical noise in non-speech regions. Therefore, first, we define kurtosis matching for DNN-based low-musical-noise speech enhancement. Kurtosis is the fourth-order moment and is known to correlate with the amount of musical noise. The kurtosis matching is a penalty term of the DNN training and works to reduce the amount of musical noise. We further extend this scheme to standardized-moment matching. The extended scheme involves using moments whose orders are higher than kurtosis and generalizes the conventional musical-noise-free method based on kurtosis matching. We formulate standardized-moment matching and explore how effectively the higher-order moments reduce the amount of musical noise. Experimental evaluation results 1) demonstrate that kurtosis matching can reduce musical noise without negatively affecting noise suppression and 2) newly reveal that the sixth-moment matching also achieves low-musical-noise speech enhancement as well as kurtosis matching.

key words: *speech enhancement, musical noise, kurtosis, moment matching, deep learning*

1. Introduction

Estimating clean speech signals from noisy ones is very important for speech-based applications [1]–[3], and speech enhancement takes a role in the front-end of the applications [4]–[9]. The back-end processes can be classified into two cases: machine and human. Speech enhancement for the former case (e.g., automatic speech recognition) should optimize objective measures such as recognition accuracy. Speech enhancement for the latter case (e.g., speech telecommunication) should improve subjective measures such as auditory impressions. Although speech enhancement techniques for machine-oriented applications have been widely studied [10], techniques that can achieve better auditory impressions have not been established yet. Also, speech enhancement for human-oriented applications is needed in portable devices because speech telecommunications often require high portability. One major approach is multi-channel speech enhancement [11], which typically involves using a microphone array to model the spatial relation between microphones and a speaker to estimate a clean speech signal. However, it is unsuitable for portable devices because the microphone array requires a large physical space. Therefore, this paper addresses how to improve the auditory impression of single-channel speech enhancement.

Deep neural network (DNN)-based speech enhancement [4]–[8] involves using a DNN to estimate clean speech from noisy speech based on a supervised-learning framework. This can achieve better noise suppression thanks to the non-linearity of DNNs. The major approach in DNN-based speech enhancement is based on a soft mask. The DNN in this approach outputs a soft mask from a noisy signal, and the enhanced signal is estimated by masking the amplitude spectrogram of the noisy signal. The DNN parameters are often optimized by minimizing an objective measure (e.g., the $L_{1,1}$ norm between amplitude spectrograms of target clean and enhanced signals). Although the trained DNN can significantly reduce noise in an observed speech signal, it often causes artifacts in the non-speech regions and degrades the auditory impressions, as shown in the left half of Fig. 1. A well-known example of such artifacts is *musical noise* [12], [13] in non-speech regions, which is artificial distortion caused by nonlinear signal pro-

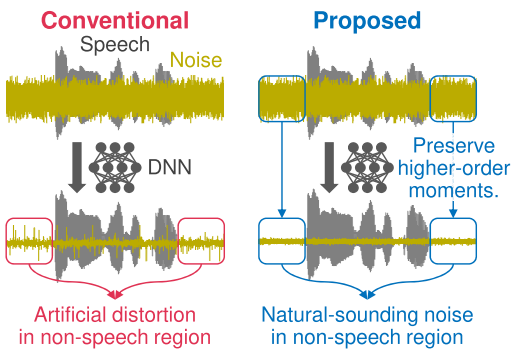

Fig. 1 Comparison of conventional and proposed DNN-based speech enhancement methods. The conventional method can significantly suppress noise in the observed signal but causes artifacts in the non-speech regions. Our method reduces the artifacts by preserving higher-order moments in the non-speech regions before and after speech enhancement.

Manuscript received February 25, 2021.
Manuscript revised June 15, 2021.
Manuscript publicized July 30, 2021.

[†]The authors are with the University of Tokyo, Tokyo, 113–8656 Japan.
a) E-mail: satoshi.mizoguchi33@gmail.com
b) E-mail: yuuki_saito@ipc.i.u-tokyo.ac.jp
c) E-mail: shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
d) E-mail: hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

DOI: 10.1587/transinf.2021EDP7041

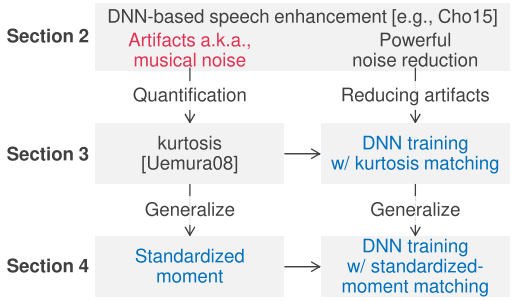

Fig. 2 The remainder of this paper.

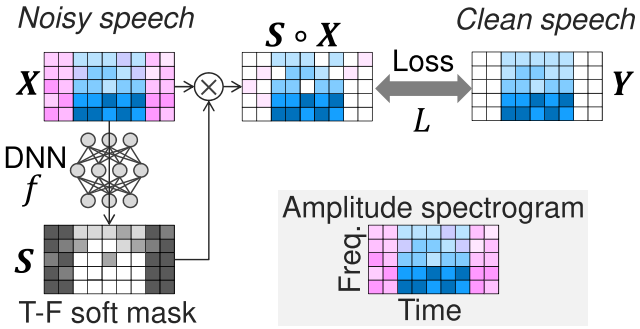

Fig. 3 Procedure of conventional DNN-based speech enhancement using a soft mask.

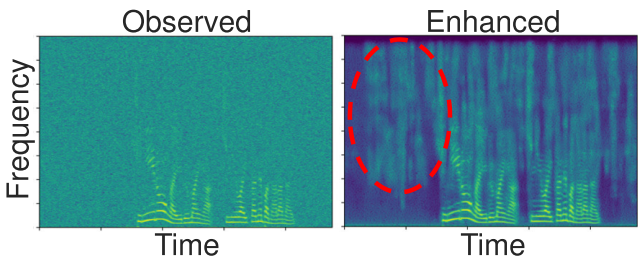

Fig. 4 Example of musical noise. The left and right are amplitude spectrograms of observed and enhanced signals, respectively. The red-circled speckled spectrogram in right figure is musical noise, which sounds artificial and significantly degrades audio impressions.

cessing (e.g., DNN-based speech enhancement).

To achieve low-musical-noise speech enhancement, we propose DNN-based speech enhancement using moment matching. First, we define *kurtosis matching* as a penalty term of the DNN training objective. *Kurtosis* is the fourth-order moment known to strongly correlate with the amount of musical noise [14]. The kurtosis matching works to prevent the increase of kurtosis in non-speech regions, as shown in the right half of Fig. 1. We further extend this scheme to *standardized-moment matching*. The extended scheme generalizes kurtosis matching and involves using moments whose orders are higher than the kurtosis. We formulate standardized-moment matching and explore how effectively high-order moments reduce musical noise, which was not fully investigated in previous speech enhancement studies. Our experimental evaluation results 1) demonstrate that kurtosis matching can reduce musical noise without negatively affecting noise suppression and 2) newly reveal that the sixth-moment matching also achieves low-musical-noise speech enhancement as well as kurtosis matching.

The remainder of this paper is organized as follows (also shown in Fig. 2). Section 2 reviews the conventional DNN-based speech enhancement method and how it degrades auditory impressions due to musical noise. Section 3 describes our proposed method using kurtosis matching, and Sect. 4 extends our method to standardized-moment matching. Section 5 provides experimental evaluation. Section 6 concludes this paper.

2. Conventional DNN-Based Speech Enhancement

This section describes conventional DNN-based speech enhancement using a soft mask. Figure 3 shows the procedure.

2.1 Mask Estimation by DNN

Let $\boldsymbol{X}$ be an amplitude spectrogram of an observed signal, which is calculated through short-term Fourier transform (STFT). A DNN $f(\cdot)$, whose model parameters are defined as $\boldsymbol{\Theta}$, outputs a soft mask $\boldsymbol{S}$ from $\boldsymbol{X}$ (i.e., $\boldsymbol{S} = f(\boldsymbol{X}; \boldsymbol{\Theta})$). The enhanced signal's amplitude spectrogram $\boldsymbol{Z}$ is calculated as the Hadamard product (i.e., element-wise product) of the input spectrogram and the soft mask (i.e., $\boldsymbol{Z} = \boldsymbol{S} \circ \boldsymbol{X}$). The $\boldsymbol{X}, \boldsymbol{S}$, and $\boldsymbol{Z}$ are the $(K+1)$-by-T matrices, and their components are defined as $X_{k,t}, S_{k,t}$ and $Z_{k,t}$, respectively. The $k \in \mathcal{K} := \{0, \cdots, K\}$ represents the frequency bin, and the $t \in \mathcal{T} := \{1, \cdots, T\}$ represents the frame index.

Given a clean signal's target amplitude spectrogram $\boldsymbol{Y}$, the DNN training's objective function is defined as

$$L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Theta}) := \|\boldsymbol{S} \circ \boldsymbol{X} - \boldsymbol{Y}\|_{1,1}, \tag{1}$$

where $\|\cdot\|_{1,1}$ is the $L_{1,1}$ norm that indicates the sum of the absolute values of each element. The DNN model parameters (i.e., weight matrices and bias vectors) are optimized to minimize the sample expectation of Eq. (1) over training data calculated as

$$\begin{aligned}\hat{\boldsymbol{\Theta}} &= \operatorname*{argmin}_{\boldsymbol{\Theta}} \mathbb{E}[L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Theta})] \\ &\simeq \operatorname*{argmin}_{\boldsymbol{\Theta}} \frac{1}{M} \sum_{m=1}^{M} L(\boldsymbol{X}_m, \boldsymbol{Y}_m; \boldsymbol{\Theta}),\end{aligned} \tag{2}$$

where $\boldsymbol{X}_m$ and $\boldsymbol{Y}_m$ denote the mth training examples. The trained DNN outputs a soft mask to extract speech components from an observed spectrogram and suppresses noise. The enhanced signal in the time domain is calculated through inverse STFT using the masked amplitude spectrogram $\boldsymbol{S} \circ \boldsymbol{X}$ and the observed signal's original phase spectrogram.

2.2 Problem: Musical Noise

Non-linear signal processing (e.g., masking amplitudes) causes musical noise in non-speech regions. Figure 4 shows an example of musical noise. The speckled spectrogram in the non-speech regions sounds very artificial and significantly degrades auditory impressions[†]. Such speckles can be seen as *outliers* of an amplitude's distribution.

3. Enhancement Using Kurtosis Matching

We define kurtosis matching in Sect. 3.1 and propose DNN-based speech enhancement using the kurtosis matching in Sect. 3.2. Figure 5 shows the DNN training procedure.

3.1 Kurtosis Definition

Kurtosis is defined as the fourth-order moment. It can evaluate the weight of a probability density function's skirts. Therefore, it can quantify outliers of the observed data. The amount of musical noise in non-speech regions is known to strongly correlate with kurtosis [14]. Let W be a nonnegative scalar random variable that follows a probability distribution $p(w)$. The nth moment of W is defined as

$$\mu_n := \int_0^\infty w^n p(w) \mathrm{d}w, \tag{3}$$

and the kurtosis of W is defined by using the second and fourth moments as

$$K_W := \frac{\mu_4}{\mu_2^2}. \tag{4}$$

Note that the original kurtosis definition in statistics is the central moment; in other words, the moment regarding the random variable's mean. However, since the zero-mean (i.e., non-central) second-order moment corresponds to noise power in a speech signal, zero-mean kurtosis is appropriate. Therefore, we define Eq. (4) as zero-mean kurtosis. Given T observed data $W_1, \cdots, W_T$, the sample kurtosis κ_W is derived by the Monte Carlo integration as

$$\kappa_W = \frac{1}{T} \frac{\sum_{t=1}^T W_t^4}{\left(\sum_{t=1}^T W_t^2\right)^2}. \tag{5}$$

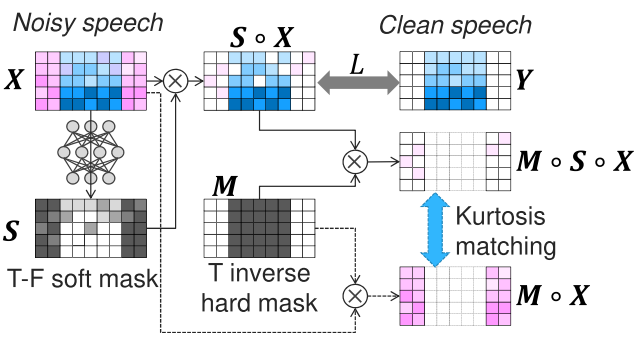

Fig. 5 Procedure of proposed DNN-based speech enhancement with kurtosis matching. The matrix M is a mask to determine the non-speech regions, such that all the elements corresponding to the non-speech regions are 1 and all other elements are 0.

[†]The speckle is also observed in the speech regions, but its effect on auditory impressions is negligible.

3.2 Kurtosis Discrepancy

We define the *kurtosis discrepancy (KD)* to quantify the increase of the kurtosis by DNN-based speech enhancement. The KD definition is inspired by the generative moment matching network (GMMN) [15], but our definition differs in that it does not use a kernel function to calculate the discrepancy. We use the frequency sub-band KD of an amplitude spectrogram in non-speech regions.

We split a set of frequency indices into $\mathcal{K}_i := \{k_i, \cdots, k_{i+1}-1\}$, where $i = 1, \cdots, N-1$, $k_1 = 0$, $k_N = K+1$. A non-speech region's set of frame indices is denoted as $\mathcal{T}' \subset \mathcal{T}$. The KD in the non-speech regions is defined as

$$\mathrm{KD}(\boldsymbol{X}, \boldsymbol{Z}) := \sum_{i=1}^N \alpha_i \left| \mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(X_{k,t}) - \mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(Z_{k,t}) \right|, \tag{6}$$

where $\mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(X_{k,t})$ is the sample kurtosis of the non-speech regions' amplitude spectrogram within the kth sub-bund, which is calculated by Eq. (5). The $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N]$ is the weight parameter of the sample kurtosis for each frequency sub-band.

The KD shown in Eq. (6) evaluates the increase of the kurtosis, but the KD value depends on the absolute value of the increase. Therefore, DNN training using only the KD term underestimates the kurtosis increase of low-kurtosis noise (e.g., Gaussian noise). We confirmed this phenomenon in our preliminary experiment. Therefore, we define yet another discrepancy called *scaled kurtosis discrepancy (SKD)* as

$$\begin{aligned}\mathrm{SKD}(\boldsymbol{X}, \boldsymbol{Z}) &:= \sum_{i=1}^N \alpha_i \left| \frac{\mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(X_{k,t}) - \mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(Z_{k,t})}{\mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(X_{k,t})} \right| \\ &= \sum_{i=1}^N \alpha_i \left| 1 - \frac{\mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(Z_{k,t})}{\mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(X_{k,t})} \right|, \end{aligned} \tag{7}$$

which corresponds to the distance between 1 and the kurtosis ratio of enhanced and observed amplitude spectrograms [14].

3.3 DNN Training

We propose DNN training for speech enhancement that considers the SKD. The SKD term shown in Eq. (7) is added to the objective function as the penalty term, and it works to reduce the amount of musical noise. The objective function is reformulated as

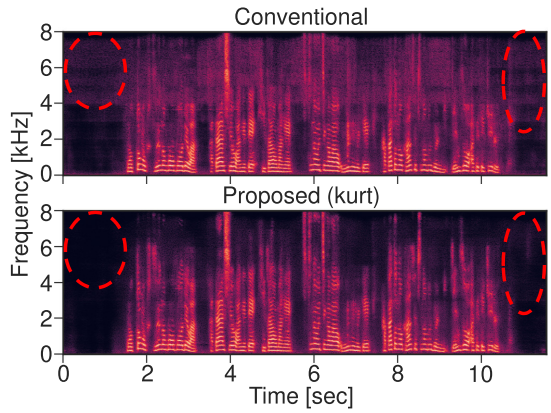

Fig. 6 Enhanced amplitude spectrograms using the conventional method (top) and the proposed kurtosis matching method (bottom). The red-circled parts in the spectrogram enhanced by the conventional method represent non-speech regions containing musical noise (i.e., speckled artifacts). The proposed method successfully removes these artifacts and achieves musical-noise-free speech enhancement.

$$L_{\mathrm{SKD}}(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta}) := L(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta}) + \lambda \mathrm{SKD}(\boldsymbol{X}, f(\boldsymbol{X};\boldsymbol{\Theta})\circ\boldsymbol{X}), \tag{8}$$

where λ is a hyperparameter to control the weight of the kurtosis matching.

Figure 6 shows amplitude spectrograms enhanced by the conventional method (i.e., without kurtosis matching) and our proposed method with kurtosis matching. We can see that speckled artifacts (i.e., musical noise) are removed by our method.

4. Enhancement Using Standardized-Moment Matching

We theoretically extend DNN training with the kurtosis matching proposed in Sect. 3. Although the kurtosis ratio is known to have a strong correlation with the amount of musical noise [14], it contains arbitrariness as an objective measure. Therefore, we propose a framework that uses higher-order moments than kurtosis does, and we explore a new method that successfully achieves low-musical-noise speech enhancement. First, we generalize kurtosis to a standardized moment in Sect. 4.1, then we define a new DNN training objective in Sect. 4.2.

4.1 Standardized Moment [16]

We define the nth standardized moment, which generalizes the kurtosis definition shown in Eq. (5) as

$$K_W^{(n)} := \frac{\mu_n}{\mu_2^{n/2}}, \tag{9}$$

where the fourth standardized moment corresponds to kurtosis (i.e., $\kappa_W^{(4)} = \kappa_W$). Since W is a non-negative random variable (i.e., an element of an amplitude spectrogram), $\kappa_W^{(n)}$ represents the amount of outliers when $n > 2$. Therefore, the use of the higher-order standardized moments should quantify the amount of musical noise. Given T observed data $W_1, \cdots, W_T$, the sample standardized moment $\kappa_W^{(n)}$ is derived by the Monte Carlo integration as

$$\kappa_W^{(n)} = \frac{\frac{1}{T}\sum_{t=1}^{T} W_t^n}{\left(\frac{1}{T}\sum_{t=1}^{T} W_t^2\right)^{n/2}} = T^{n/2-1}\frac{\sum_{t=1}^{T} W_t^n}{\left(\sum_{t=1}^{T} W_t^2\right)^{n/2}}. \tag{10}$$

4.2 DNN Training with Standardized-Moment Matching

Similar to the kurtosis ratio [14], we define the *standardized-moment ratio* as

$$\frac{\kappa_{\hat{\boldsymbol{S}}\circ\boldsymbol{X}}^{(n)}}{\kappa_{\boldsymbol{X}}^{(n)}}. \tag{11}$$

The ratio can quantify the increase of the standardized moment due to the soft-masking process. The nth standardized moment of the amplitude spectrogram $\boldsymbol{X}$ in non-speech regions is calculated as

$$\underset{\mathcal{K}_i,\mathcal{T}';n}{\mathrm{SM}}(\boldsymbol{X}) = \frac{T'^{n/2-1}\sum_{t\in\mathcal{T}'}\sum_{k\in\mathcal{K}_i} X_{k,t}^n}{\left(\sum_{t\in\mathcal{T}'}\sum_{k\in\mathcal{K}_i} X_{k,t}^2\right)^{n/2}}, \tag{12}$$

where T' denotes the number of frames in the non-speech regions. We define the nth-order scaled standardized-moment discrepancy (n-SSMD) as

$$\begin{aligned}\mathrm{SSMD}^{(n)}(\boldsymbol{X},\boldsymbol{Z}) &:= \sum_{i=1}^{N}\alpha_i\left|\frac{\mathrm{SM}_{\mathcal{K}_i,\mathcal{T}';n}(\boldsymbol{X}) - \mathrm{SM}_{\mathcal{K}_i,\mathcal{T}';n}(\boldsymbol{Z})}{\mathrm{SM}_{\mathcal{K}_i,\mathcal{T}';n}(\boldsymbol{X})}\right| \\ &= \sum_{i=1}^{N}\alpha_i\left|1 - \frac{\mathrm{SM}_{\mathcal{K}_i,\mathcal{T}';n}(\boldsymbol{Z})}{\mathrm{SM}_{\mathcal{K}_i,\mathcal{T}';n}(\boldsymbol{X})}\right|,\end{aligned} \tag{13}$$

which corresponds to the distance between 1 and Eq. (11). Therefore, the kurtosis matching is the special case of n-SSMD (i.e., 4-SSMD is equivalent to Eq. (7)).

We propose a training objective for DNN-based speech enhancement that considers the increase of some standardized moments as the penalty terms. Given the set of moment orders to be considered $\mathcal{N}$, we define the loss function as

$$L_{\mathrm{SSMD}}(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta}) := L(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta}) + \lambda\sum_{n\in\mathcal{N}}\gamma_n \mathrm{SSMD}^{(n)}(\boldsymbol{X}, f(\boldsymbol{X};\boldsymbol{\Theta})\circ\boldsymbol{X}), \tag{14}$$

where γ_n is a hyperparameter that determines the relative importance of each n-SSMD term, which satisfies $\sum_{n\in\mathcal{N}}\gamma_n = 1$. By performing speech enhancement using DNNs trained to minimize Eq. (14) with various settings of $\mathcal{N}$ and each γ_n, and evaluating the results subjectively, we can explore higher-order standardized moments other than kurtosis that can reduce the amount of musical noise.

5. Experimental Evaluation

5.1 Experimental Conditions

The training and evaluation data were 31,869 utterances selected from the JNAS corpus [17] and 200 utterances selected from the JSUT corpus [18]. These corpora were Japanese clean speech corpora, and the sentences and speakers did not overlap between the corpora. Fixed-length silence regions were concatenated to the utterances. Six kinds of noise were prepared, and they were artificially added to the silence-concatenated utterances with $\{-5, 0, 5, 10\}$ dB of the SNR settings. They consisted of five real noise recordings selected from the DEMAND corpus [19] (PSTATION, PRESTO, NFIELD, SPSQUARE, and TBUS), and one artificial Gaussian noise (GAUSS). Table 1 lists the six kinds of noise we used in this evaluation to cover a wide range of kurtosis. Note that we did not build noise-specific enhancement models, i.e., we trained single DNN by using noisy speech generated with these six noise cases and four SNR settings. The sampling rate was 16 kHz. The window function of STFT was the 1,024-tap Hanning window. The hop size of STFT was 80. The DNN architecture was U-Net [20], which consisted of 12 hidden convolutional layers, Leaky ReLU activation [21], and Dropout [22]. The minibatch size was set to 32, and the patch length was set to 256. The number of sub-bands N was set to 4. The sub-band indices were set to $\kappa_1 = [0, \cdots, 127], \kappa_2 = [128, \cdots, 255], \kappa_3 = [256, \cdots, 383]$, and $\kappa_4 = [384, \cdots, 512]$. Weights of the SKD terms for each sub-band were empirically set to $[0.01, 1.0, 1.0, 1.0]$. A weight of the kurtosis matching term or standardized-moment matching, i.e., λ in Eq. (8) or Eq. (14), was empirically set to 1×10^{-4}. An Adam optimizer [23] with 0.01 of the learning rate was used for the DNN training. The number of training iterations was 30.

5.2 Conventional Method vs. Proposed Kurtosis Matching

This section confirms that our proposed kurtosis matching can reduce musical noise without negatively affecting noise suppression. The methods to be compared are labeled as follows.

- **Conventional**: conventional DNN-based speech enhancement (Sect. 2)
- **Proposed (kurt)**: proposed method with kurtosis matching (Sect. 3)

Table 1 Noise and its kurtosis used for experimental evaluation

Noise	Description	Kurtosis
GAUSS	Gaussian noise	3.00
PSTATION	Busy subway station	5.56
PRESTO	University restaurant	12.1
NFIELD	Sports field	13.3
SPSQUARE	Public town square	29.8
TBUS	Public transit bus	35.8

5.2.1 Objective Evaluation

We calculated the following objective evaluation metrics of enhanced signals.

- **Signal-to-distortion ratio (SDR) improvement**: a basic criterion to measure speech enhancement performance. Higher is better.
- **Cepstrum distortion (CD) in speech regions**: distortion in speech region. Smaller is better.
- **Kurtosis ratio (KR) in non-speech regions**: 1 means that speech enhancement does not change kurtosis.
- **Perceptual evaluation of speech quality (PESQ) [24]**: a commonly used criterion to evaluate speech quality of enhanced speech objectively. Higher is better.

These values were calculated over all the evaluation utterances.

Table 2 lists the median values of the four evaluation metrics (see Appendix A for their box plots). In all the settings of the SNR and noise, "Proposed (kurt)" slightly deteriorates SDR improvement, CD, and PESQ. This is reasonable because the proposed SKD term works as the regularization during training and does not necessarily improve these criteria more than the conventional method trained to minimize Eq. (1) (i.e., the $L_{1,1}$ norm) only. However, the degrees of deterioration can be acceptable ($<$ 3 dB for SDR improvement, $<$ 1.5 dB for CD, and $<$ 0.43 for PESQ), and we can expect the negative effects of the proposed kurtosis matching on noise suppression to be very small. Focusing on the KR value results, "Proposed (kurt)" significantly decreases the values in all the SNR and noise settings compared with "Conventional." Note that the KR values of "Proposed (kurt)" does not match with 1 because the method does not theoretically guarantee to fix kurtosis before and after speech enhancement. These results suggest that our proposed method with kurtosis matching reduces musical noise with little degradation of noise suppression.

5.2.2 Subjective Evaluation

We subjectively evaluated the audio impressions of non-speech regions after utilizing the conventional and proposed DNN-based speech enhancement methods. The non-speech regions used for this evaluation were trimmed from enhanced speech. Preference AB tests (listening tests) on naturalness were conducted in our crowdsourcing evaluation system. We used the crowdsourcing platform "Lancers" [25]. We presented listeners with pairs of the non-speech regions of the conventional and proposed methods in random order. Each listener selected the one that sounded more natural. The listening tests were conducted for each of the SNR and noise settings. Twenty-four listeners participated in each test. Each listener answered for ten pairs, and 240 answers were collected for each test. The total number of listeners was 6 (noise settings) $\times$ 4 (SNR settings) $\times$ 24 (listeners) = 576.

Table 2 Median values of SDR improvement, CD, KR, and PESQ. The values were calculated over all the evaluation data. Evaluation results of conventional method ("Conventional"), proposed methods with kurtosis matching ("Proposed (kurt)"), and standardized-moment matching ("Proposed (6th)") are listed

		SDR improvement [dB]			CD [dB]			KR			PESQ		
Noise	SNR	Conventional	Proposed (kurt)	Proposed (6th)	Conventional	Proposed (kurt)	Proposed (6th)	Conventional	Proposed (kurt)	Proposed (6th)	Conventional	Proposed (kurt)	Proposed (6th)
GAUSS	-5 dB	19.53	16.85	16.69	11.65	12.77	12.37	138.07	7.25	4.52	1.89	1.64	1.56
	0 dB	17.22	15.42	15.29	10.98	12.63	11.98	223.13	7.79	5.19	2.15	1.89	1.79
	5 dB	14.88	13.59	13.62	10.09	12.02	10.78	290.58	8.62	6.04	2.40	2.16	2.01
	10 dB	12.66	11.52	11.63	9.02	10.87	9.44	296.59	8.79	6.06	2.66	2.42	2.23
PSTATION	-5 dB	17.83	16.28	16.24	8.43	9.42	9.05	60.09	3.16	3.02	1.84	1.74	1.62
	0 dB	16.61	14.64	14.70	6.03	7.14	6.96	64.38	7.02	5.30	2.30	2.07	1.92
	5 dB	15.02	12.71	12.75	4.42	5.37	5.14	58.30	11.82	8.15	2.62	2.34	2.20
	10 dB	13.36	10.66	10.75	3.35	4.03	3.78	47.41	13.37	8.51	2.90	2.61	2.49
PRESTO	-5 dB	10.05	10.30	9.40	10.51	11.32	10.95	12.44	2.94	4.31	1.26	1.27	1.20
	0 dB	10.94	10.92	10.66	9.25	10.40	10.02	11.41	2.80	3.52	1.73	1.65	1.55
	5 dB	10.18	10.14	10.06	7.85	8.91	8.42	11.54	2.13	2.15	2.15	1.99	1.85
	10 dB	8.97	8.67	8.64	5.54	6.74	6.26	11.59	2.33	1.73	2.51	2.30	2.14
NFIELD	-5 dB	25.84	22.78	22.99	5.18	5.90	4.78	23.36	4.69	3.73	2.63	2.43	2.29
	0 dB	23.57	20.36	20.57	3.90	4.31	3.99	20.04	4.94	4.31	2.96	2.74	2.61
	5 dB	21.29	17.69	17.86	3.12	3.58	3.44	19.40	5.97	5.85	3.22	2.99	2.90
	10 dB	18.42	14.21	14.54	2.50	2.75	2.85	19.22	6.42	6.58	3.47	3.27	3.21
SPSQUARE	-5 dB	18.27	17.09	17.08	5.86	6.02	5.00	8.63	3.02	2.52	1.97	1.84	1.71
	0 dB	18.05	16.58	16.54	4.34	4.72	4.26	8.23	3.01	2.68	2.41	2.23	2.09
	5 dB	17.31	15.36	15.30	3.35	3.80	3.69	9.31	2.62	2.50	2.77	2.58	2.44
	10 dB	15.84	13.52	13.48	2.76	3.18	3.26	10.48	3.34	2.88	3.05	2.88	2.77
TBUS	-5 dB	26.21	23.27	23.30	4.21	4.58	4.15	8.25	4.23	3.26	2.72	2.55	2.40
	0 dB	23.97	21.25	21.25	3.32	3.67	3.64	6.94	4.38	3.72	3.02	2.85	2.72
	5 dB	21.89	18.96	19.17	2.56	2.83	2.94	7.79	4.66	4.66	3.32	3.17	3.09
	10 dB	18.88	15.46	15.90	2.29	2.41	2.51	9.13	5.27	4.95	3.55	3.44	3.39

Table 3 lists the result. "Proposed (kurt)" significantly outperforms "Conventional" in almost every case except for the "GAUSS" case. We predict that the quality deterioration in the "GAUSS" case will not be a serious problem in practice since the Gaussian noise is artificially generated. These results demonstrate that our proposed method with kurtosis matching can reduce musical noise in DNN-based speech enhancement to achieve better auditory impressions.

5.3 Kurtosis Matching vs. Standardized-Moment Matching

We investigated the effectiveness of higher-order standardized moments in musical noise reduction by using the proposed standardized-moment matching. Here, we used both kurtosis and the sixth-order moment, with various weight parameter settings $\gamma_6 = [0.00, 0.25, 0.50, 0.75, 1.00]$ and $\gamma_4 = 1 - \gamma_6$. The setting with $\gamma_6 = 0.0$ is equivalent to "Proposed (kurt)" (i.e., using only kurtosis), and that with $\gamma_6 = 1.0$ is equivalent to using only the sixth-order moment. We labeled the latter case "Proposed (6th)."

5.3.1 Objective Evaluation

As in Sect. 5.2.1, we calculated SDR improvement, CD, KR, and PESQ. Table 2 lists the results of $\gamma_6 = 0.0$ ("Proposed (kurt)") and $\gamma_6 = 1.0$ ("Proposed (6th)") settings. No significant differences between the two methods are observed among the SDR improvement, CD, and PESQ. Curiously, "Proposed (6th)" improves the KR values more than "Proposed (kurt)" does using some settings; although it was out of the consideration during training. This result newly reveals that the sixth-moment matching can also reduce the amount of musical noise as a side effect similar to kurtosis matching. Note that the use of intermediate settings (i.e.,

Table 3 Preference scores on naturalness of noise with ξ^2-test's p-values. Conventional method ("Conventional") and Proposed method with kurtosis matching ("Proposed (kurt)") are compared. **Bold** indicates significantly (p-value < 0.05) better scores

Noise label	Input SNR	Conventional	Proposed (kurt)	p-value
GAUSS	−5 dB	**0.688**	0.313	$< 10^{-10}$
	0 dB	**0.804**	0.196	$< 10^{-10}$
	5 dB	**0.725**	0.275	$< 10^{-10}$
	10 dB	**0.863**	0.138	$< 10^{-10}$
PSTATION	−5 dB	0.408	**0.592**	5.34×10^{-5}
	0 dB	0.417	**0.583**	2.45×10^{-4}
	5 dB	0.404	**0.596**	2.36×10^{-5}
	10 dB	0.271	**0.729**	$< 10^{-10}$
PRESTO	−5 dB	0.517	0.483	4.66×10^{-1}
	0 dB	0.421	**0.579**	4.98×10^{-4}
	5 dB	0.354	**0.646**	$< 10^{-10}$
	10 dB	0.483	0.517	4.66×10^{-1}
NFIELD	−5 dB	0.250	**0.750**	$< 10^{-10}$
	0 dB	0.221	**0.779**	$< 10^{-10}$
	5 dB	0.200	**0.800**	$< 10^{-10}$
	10 dB	0.192	**0.808**	$< 10^{-10}$
SPSQUARE	−5 dB	0.367	**0.633**	2.94×10^{-9}
	0 dB	0.383	**0.617**	2.34×10^{-7}
	5 dB	0.225	**0.775**	$< 10^{-10}$
	10 dB	0.250	**0.750**	$< 10^{-10}$
TBUS	−5 dB	0.238	**0.763**	$< 10^{-10}$
	0 dB	0.258	**0.742**	$< 10^{-10}$
	5 dB	0.167	**0.833**	$< 10^{-10}$
	10 dB	0.238	**0.763**	$< 10^{-10}$

Table 4 Preference scores on naturalness of noise with ξ^2-test's p-values. Proposed methods with kurtosis matching ("Proposed (kurt)") and standardized-moment matching ("Proposed (6th)") are compared. **Bold** indicates significantly (p-value < 0.05) better scores

Noise label	Input SNR	Proposed (kurt)	Proposed (6th)	p-value
GAUSS	−5 dB	0.508	0.492	8.00×10^{-1}
	0 dB	0.456	0.544	1.64×10^{-1}
	5 dB	0.560	0.440	5.78×10^{-2}
	10 dB	0.476	0.524	4.48×10^{-1}
PSTATION	−5 dB	0.432	0.568	3.15×10^{-2}
	0 dB	0.464	0.536	2.55×10^{-1}
	5 dB	0.444	0.556	7.66×10^{-2}
	10 dB	0.524	0.476	4.48×10^{-1}
PRESTO	−5 dB	0.444	0.556	7.66×10^{-2}
	0 dB	0.484	0.516	6.13×10^{-1}
	5 dB	0.504	0.496	8.99×10^{-1}
	10 dB	0.516	0.484	6.13×10^{-1}
NFIELD	−5 dB	0.424	0.576	1.62×10^{-2}
	0 dB	0.428	0.572	2.28×10^{-2}
	5 dB	0.336	**0.664**	2.15×10^{-7}
	10 dB	0.312	**0.688**	2.76×10^{-9}
SPSQUARE	−5 dB	0.544	0.456	1.64×10^{-1}
	0 dB	0.544	0.456	1.64×10^{-1}
	5 dB	0.564	0.436	4.30×10^{-2}
	10 dB	0.476	0.524	4.48×10^{-1}
TBUS	−5 dB	**0.648**	0.352	2.87×10^{-6}
	0 dB	**0.612**	0.388	3.97×10^{-4}
	5 dB	**0.620**	0.380	1.48×10^{-4}
	10 dB	**0.644**	0.356	5.27×10^{-6}

$\gamma_6 = \{0.25, 0.50, 0.75\}$) does not show any improvements (see Appendix B), and the following subjective evaluation compares only "Proposed (kurt)" and "Proposed (6th)."

5.3.2 Subjective Evaluation

As in Sect. 5.2.2, the preference AB tests were conducted on the naturalness of non-speech regions processed by "Prop (kurt)" and "Prop. (6th)." Table 4 lists the result. The scores of "Proposed (6th)" are comparable or better than those of "Proposed (kurt)" in all cases except for "TBUS" noise. This result indicates that we can use both kurtosis and the sixth-order moment to achieve low-musical-noise speech enhancement. One possible reason for the quality degradation in the "TBUS" case is unstable modeling due to the highest kurtosis among the six noise settings, which increases the ratio of outliers and makes the training based on standardized-moment matching more difficult.

6. Conclusion

We proposed DNN-based speech enhancement with kurtosis or standardized-moment matching. The scaled discrepancy between kurtosis or the standardized moment of enhanced and observed amplitude spectrograms is introduced as the penalty term of DNN training, which works to reduce musical noise in non-speech regions. The experimental evaluation results 1) demonstrated that DNN training with kurtosis matching reduced musical noise and 2) newly revealed that the sixth-moment matching also achieved low-musical-noise speech enhancement as well as kurtosis matching. Our future work includes a detailed investigation of the hyperparameter settings of the proposed method.

Acknowledgements

Part of this work was supported by JSPS KAKENHI 19H01116.

References

[1] K. Kobayashi, Y. Haneda, K. Furuya, and A. Kataoka, "A hands-free unit with noise reduction by using adaptive beamformer," IEEE Trans. Consum. Electron., vol.54, no.1, pp.116–122, Feb. 2008.
[2] Y. Hioka, K. Furuya, K. Kobayashi, S. Sakauchi, and Y. Haneda, "Angular region-wise speech enhancement for hands-free speakerphone," IEEE Trans. Consum. Electron., vol.58, no.4, pp.1403–1410, Nov. 2012.
[3] T. Traphagan, J.V. Kucsera, and K. Kishi, "Impact of class lecture webcasting on attendance and learning," Educational Technology Research and Development, vol.58, no.1, pp.19–37, Feb. 2010.
[4] 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 and Language Processing, vol.23, no.1, pp.7–19, Jan. 2015.
[5] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," Proc. INTERSPEECH, pp.3678–3772, San Francisco, U.S.A., Sept. 2016.
[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J.R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," Proc. 12th Int. Conf. Latent Variable Analysis and Signal Separation, vol.9237, pp.91–99, Liberec, Czech Republic, Aug. 2015.
[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," Proc. INTERSPEECH, pp.436–440, Lyon, France, Aug. 2013.
[8] S. Leglaive, U. Simsekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," IEEE Int. Conf. Acoust., Speech, Signal Process., pp.541–545, Brighton, United Kingdom, May 2019.
[9] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements," Proc. Int. Conf. Acoust., Speech, Signal Process., pp.81–85, New Orleans, LA, U.S.A., March 2017.
[10] A.H. Moore, P.P. Parada, and P.A. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," Computer Speech and Language, vol.246, pp.574–584, Nov. 2017.
[11] T.H. Dat, K. Takeda, and F. Itakura, "Multichannel speech enhancement based on generalized gamma prior distribution with its online adaptive estimation," IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.439–447, March 2008.
[12] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.345–349, April 1994.
[13] Z. Goh, K.-C. Tan, and B. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," IEEE Trans. Speech Audio Process., vol.6, no.3, pp.287–292, May 1998.
[14] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," Proc. International Workshop for Acoustic Echo and Noise Control, Seattle, W.A., U.S.A., Sept. 2008.
[15] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," Proc. 32th Int. Conf. Machine Learning, vol.37, pp.1718–

1727, Lille, France, July 2015.
[16] J.F. Kenny and E.S. Keeping, "Moments in standard units," Mathematics of Statistics, Pt. 1, 3rd ed. pp.98–99, Princeton, NJ: Van Nostrand, 1962.
[17] "JNAS," http://research.nii.ac.jp/src/JNAS.html, accessed: 2018-12-05.
[18] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv preprint, 1711.00354, Oct. 2017.
[19] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," IEEE Trans. Audio, Speech, Language Process., vol.20, no.7, pp.2080–2094, Sept. 2021.
[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," Proc. 18th Int. Conf. Medical Image Computing and Computer Assisted Intervention, pp.234–241, Munich, Germany, Oct. 2015.
[21] A.L. Maas, A.Y. Hanuun, and A.Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," Proc. 30th Int. Conf. Machine Learning, vol.30, Atlanta, Georgia, U.S.A., June 2013.
[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol.15, pp.1929–1958, June 2014.
[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. Int. Conf. Learning Representations, Banff, Canada, Dec. 2014.
[24] 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," Proc. ICASSP, pp.749–752, Salt Lake City, U.S.A., April 2001.
[25] "Lancers," https://www.lancers.jp/.

Appendix A: Detailed Visualizations for Objective Measures

To deeply discuss the objective evaluation results shown in Table 2, we show the box plots of the SDR improvement, CD, KR, and PESQ values in Fig. A· 1, Fig. A· 2, Fig. A· 3, and Fig. A· 4, respectively. The box indicates the first, second (i.e., median), and third quartiles. An upper limit of whisker is 1.5 times longer than an interquartile range. Points denote outlier values. As one example, results of "PSTATION" noise were drawn. The median values of the criteria are different among SNR settings, but the overall tendencies of the two methods' results are similar regardless of the settings.

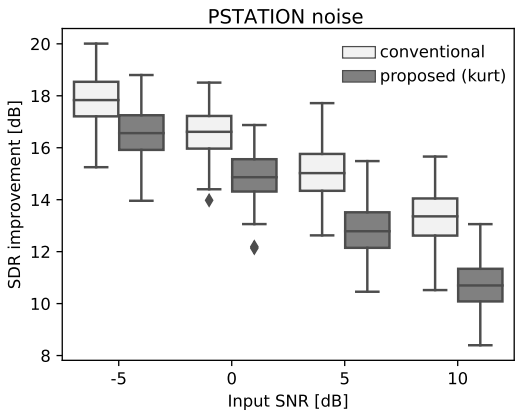

Fig. A· 1 Box plots of SDR improvement. The conventional method and proposed method with kurtosis matching are compared. Higher is better.

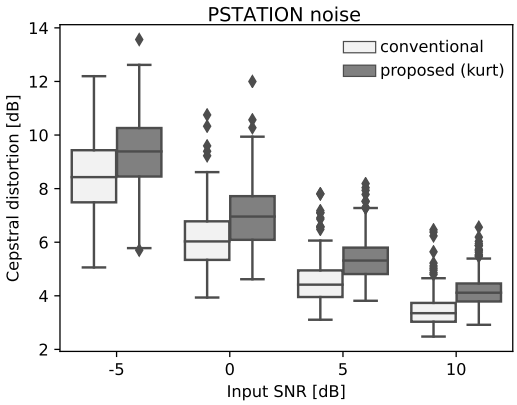

Fig. A· 2 Box plots of CD. The conventional method and proposed method with kurtosis matching are compared. Lower is better.

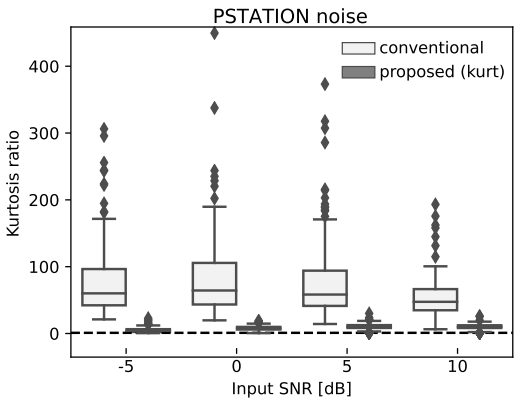

Fig. A· 3 Box plots of KR. The conventional method and proposed method with kurtosis matching are compared. 1 indicates no increase in kurtosis by speech enhancement.

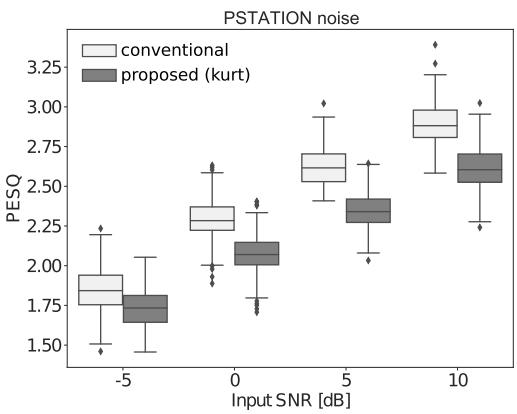

Fig. A· 4 Box plots of PESQ. The conventional method and proposed method with kurtosis matching are compared. Higher is better.

Appendix B: Investigation of Combined Kurtosis and Standardized-Moment Matching

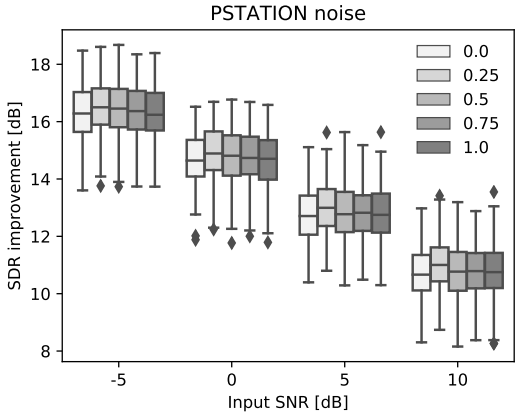

Fig. A· 5 Box plots of SDR improvement. Our standardized-moment-matching-based proposed method with several weight settings are compared. Higher is better.

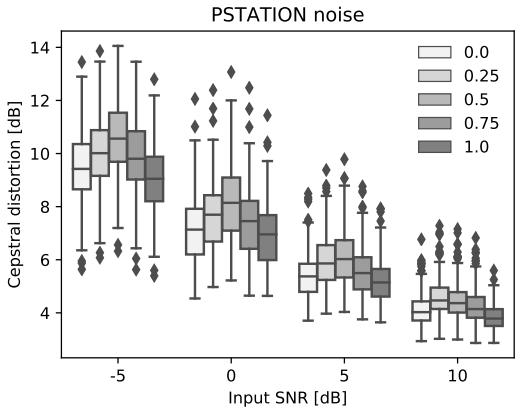

Fig. A· 6 Box plots of cepstrum distortion. Our standardized-moment-matching-based proposed method with several weight settings are compared. Lower is better.

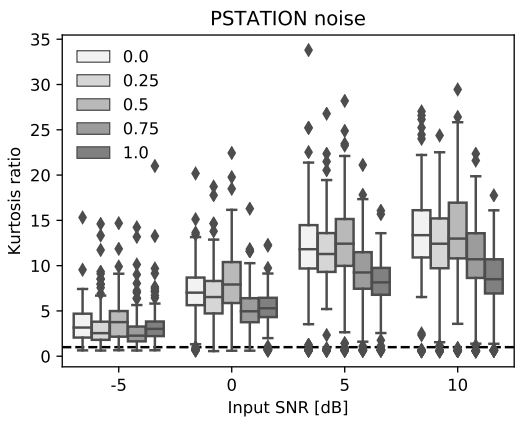

Fig. A· 7 Box plots of kurtosis ratio. Our standardized-moment-matching-based proposed method with several weight settings are compared. 1 indicates no increase in kurtosis by speech enhancement.

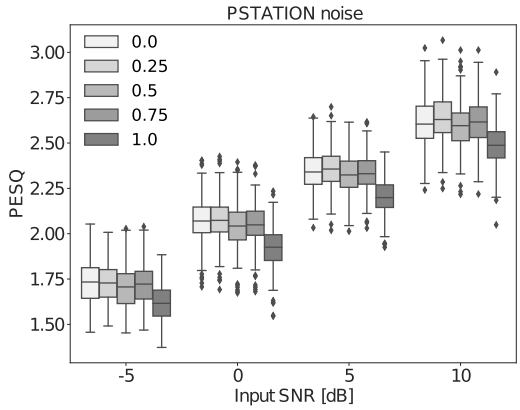

Fig. A· 8 Box plots of PESQ. Our standardized-moment-matching-based proposed method with several weight settings are compared. Higher is better.

We investigated the effects of the combinations of hyper-parameters (γ_4, γ_6) used in Sect. 5.3.1. As in Appendix A, we show the box plots of SDR improvement, CD, KR, and PESQ with various settings of γ_6 in Fig. A· 5, Fig. A· 6, Fig. A· 7, and Fig. A· 8, respectively. These results allow us to empirically determine that the intermediate settings of the fourth and sixth moments matching methods (i.e., $\gamma_6 = \{0.25, 0.50, 0.75\}$) do not benefit the objective evaluation criteria, compared with "Proposed (kurt)" (i.e., $\gamma_6 = 0.0$) and "Proposed (6th)" (i.e., $\gamma_6 = 1.0$).

Satoshi Mizoguchi received his M.S. degree from the Graduate School of Information Science and Technology, The University of Tokyo, Japan, in 2020. He received the 18th Best Student Presentation Award of ASJ. His research interests include speech enhancement, high-order statistics, and machine learning.

Yuki Saito received his M.S. degree from the Graduate School of Information Science and Technology, The University of Tokyo, Japan, in 2018. He is currently a Ph.D. student at The University of Tokyo. His research interests include speech synthesis, voice conversion, and machine learning. He has received more than ten paper awards including the 2020 IEEE Signal Processing Society Young Author Best Paper Award.

Shinnosuke Takamichi received the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016. He is currently an Assistant Professor at The University of Tokyo. He has received more than 20 paper/achievement awards including the 2020 IEEE Signal Processing Society Young Author Best Paper Award.

Hiroshi Saruwatari received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM IS Laboratory, Japan, in 1993, and Nara Institute of Science and Technology, Japan, in 2000. From 2014, he is currently a Professor of The University of Tokyo, Japan. His research interests include statistical audio signal processing, blind source separation (BSS), and speech enhancement. He has put his research into the world's first commercially available Independent-Component-Analysis based BSS microphone in 2007. He received paper awards from IEICE in 2001 and 2006, from TAF in 2004, 2009, 2012, and 2018, from IEEE IROS2005 in 2006, and from APSIPA in 2013 and 2018. He received DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, The Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hoko-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ. He is an APSIPA Distinguished Lecturer from 2018.