END-TO-END SOUND SOURCE ENHANCEMENT USING DEEP NEURAL NETWORK IN THE MODIFIED DISCRETE COSINE TRANSFORM DOMAIN

Yuma Koizumi[1]*, Noboru Harada*[1]*, Yoichi Haneda*[2]*, Yusuke Hioka*[3]*, and Kazunori Kobayashi*[1]

[1]: NTT Media Intelligence Laboratories, Tokyo, Japan

[2]: The University of Electro-Communications, Tokyo, Japan

[3]: Department of Mechanical Engineering, University of Auckland, Auckland, New Zealand

ABSTRACT

This paper presents an end-to-end deep neural network (DNN)-based source enhancement on the basis of a time-frequency (T-F) mask processing in the modified discrete cosine transform (MDCT)-domain. To retrieve the target signal perfectly in the discrete Fourier transform (DFT)-domain, both amplitude and phase of the spectrum need to be manipulated. However, since it is difficult to deal with complex values by neural network straightforward way, a real-valued T-F mask is commonly estimated and only amplitude spectrum is manipulated. In this study, we use the MDCT instead of the DFT and estimate real-valued T-F masks in the MDCT-domain. The perfect retrieval can be achieved by manipulating only the real-valued MDCT-spectra. To reduce time-domain aliasing arises from manipulating the MDCT spectrum, we build an end-to-end DNN-based source enhancement using T-F mask and train the DNN to minimize an objective function defined in the time-domain. In experiments using several kinds of objective sound quality scores, we observed that the scores were significantly improved.

***Index Terms*—** Sound source enhancement, modified discrete cosine transform (MDCT), deep learning, and end-to-end.

1. INTRODUCTION

Sound source enhancement has been studied for many years [1] because of the high demand for improving the performance of various practical applications such as automatic speech recognition [2, 3], hands-free telecommunication [4], hearing aids [5–7], and immersive audio field representation [8]. The goal of this study is to retrieve a target source from an observed signal recorded in noisy environment. To achieve this goal, a time-frequency (T-F) mask such as a Wiener filter [9] has commonly been used. To accurately estimate the T-F mask, various approaches have been developed including multi-channel [10], and non-negative matrix factorization (NMF)-based approaches [11].

A major breakthrough in T-F mask estimation has been to apply deep learning. In these approaches, a deep neural network (DNN) and/or a long short-term memory network (LSTM) have been used as a regression function to estimate a T-F mask, and signal-to-distortion ratio (SDR) has been significantly improved [12–17]. Hereafter, we call source enhancement using a T-F mask estimated by neural networks "DNN-based source enhancement." In more recent studies, neural netowrks are trained with a perceptually motivated objective function, such as perceptual weighted mean-squared-error (MSE) [18] and objective sound quality assessment scores, *e.g.*, perceptual evaluation of speech quality (PESQ) [19].

In traditional source enhancement, the estimated T-F mask is multiplied to the observed signal in the discrete-Fourier-transform (DFT)-domain. In the DFT-domain, both amplitude and phase of the spectrum have to be manipulated in order to retrieve the target signal perfectly [20]. Nevertheless, most conventional studies estimate a real-valued T-F mask and manipulate only the amplitude spectrum because it is difficult to deal with complex values by neural network straightforward way. In recent years, some attempts have been made to deal with complex value [21–23]. To manipulate both the amplitude and phase of the spectrum, a complex-valued T-F mask called a complex-ideal-ratio-mask (cIRM) is proposed [21, 22]. In this approach, neural networks are constructed to jointly estimate the real and imaginary parts of the cIRM. Thus, the number of DNN output units doubles, and an even more massive amount of training data is required to avoid over-fitting. Therefore, it might be better to use non-redundant frequency transformation, *i.e.* critically sampled transform.

In this study, we propose an end-to-end DNN-based source enhancement method in the modified-discrete-cosine-transform (MDCT)-domain [24, 25]. The key idea of this study is to use the MDCT for frequency transformation instead of the DFT. Since the MDCT is a real-valued transform, not only the amplitude spectrum but also the phase spectrum can be manipulated using real-valued T-F mask. In addition, since the MDCT is a lapped transform, the number of spectrum need to be manipulated is smaller than that of the cIRM or other real-valued transformation methods such as the discrete-cosine-transform (DCT). A problem for realizing this idea is that directly manipulating the MDCT spectrum causes time-domain aliasing [26]. To reduce the time-domain aliasing, we define an objective function that is used for training DNN in the time-domain, resulting in extending the T-F mask-based DNN source enhancement to an end-to-end system.

Section 2 introduces DNN-based source enhancement in the DFT-domain. In Section 3, we propose an end-to-end source enhancement in the MDCT-domain. After showing the SDR, the PESQ and the short-time objective intelligibility measure (STOI) [27] of the proposed method were higher than that of conventional methods in almost all input signal-to-noise ratio conditions in Section 4, we conclude this paper with some remarks in Section 5.

2. CONVENTIONAL METHOD

2.1. Sound source enhancement with T-F mask in DFT-domain

Let us consider the problem of estimating a target signal s_t, which is surrounded by an ambient noise n_t represented in the time-domain. A signal observed with a single microphone x_t is assumed to be modeled as

$$x_t = s_t + n_t, \tag{1}$$

where $t = \{1, 2, ..., T\}$ denotes the sample index in the time-domain. Here, we assume that the observed signal is split into K time-frames with overlap. Then, by applying DFT, the equation (1) can be written in the DFT-domain as

$$X_{\omega,k} = S_{\omega,k} + N_{\omega,k}, \tag{2}$$

where $X_{\omega,k}$, $S_{\omega,k}$, and $N_{\omega,k}$ are the observed, target, and noise signals in the DFT-domain, respectively. The indices $\omega = \{1, 2, ..., \Omega\}$ and $k = \{1, 2, ..., K\}$ denote the frequency and time-frame, respectively.

In sound source enhancement using T-F masks in the DFT-domain, the output signal $\hat{S}_{\omega,k}$ is obtained by multiplying a T-F mask to $X_{\omega,k}$ as

$$\hat{S}_{\omega,k} = G_{\omega,k} X_{\omega,k}, \tag{3}$$

where $G_{\omega,k}$ is a T-F mask such as the frame-wise Wiener filter [9]. Then, the output signal is transformed back to the time-domain by applying inverse-DFT (IDFT) and overlap-add. Since the unknown parameter of T-F masking is $G_{\omega,k}$, we need to estimate $G_{\omega,k}$ from $X_{\omega,k}$.

2.2. T-F mask estimation via deep learning

Various researchers have proposed applying deep learning (DL) to estimate T-F masks [12–17]. The typical DL approach estimates vectorized T-F masks for all frequency bins $\boldsymbol{G}_k := (G_{1,k}, ..., G_{\Omega,k})^\top$ as

$$\hat{\boldsymbol{G}}_k = \mathcal{M}(\boldsymbol{\phi}_k|\Theta), \tag{4}$$

where $\mathcal{M}$ is a neural network-based regression function implemented by DNN [8, 15] and/or LSTM [12, 14], $\boldsymbol{\phi}_k$ is the input acoustic feature of k-th frame extracted from the observation, Θ is the parameters of neural networks, and $\top$ denotes transposition.

In typical approaches of DNN-based source enhancement, $\mathcal{M}$ is implemented so as to estimate a real-valued T-F mask. As an implementation of the real-valued T-F mask, the phase sensitive spectrum approximation (PSA) is proposed [14]. The PSA is a real-valued T-F mask that minimizes the squared error between $S_{\omega,k}$ and $\hat{S}_{\omega,k}$ on the complex-plane. Thus, Θ is trained to minimize the following MSE on the complex-plane as

$$\mathcal{J}^{\mathrm{PSA}}(\Theta) = \sum_{k=1}^{K} ||\mathbf{S}_k - \mathcal{M}(\boldsymbol{\phi}_k|\Theta) \odot \mathbf{X}_k||_2, \tag{5}$$

where $\mathbf{S}_k := (S_{1,k}, ..., S_{\Omega,k})^\top$, $\mathbf{X}_k := (X_{1,k}, ..., X_{\Omega,k})^\top$, $|| \cdot ||_p$ is L_p norm and $\odot$ is element-wise product. The number of DNN output units is half the frame size of the DFT, *i.e.* Ω, and the activation function of the output layer of $\mathcal{M}(\boldsymbol{\phi}_k|\Theta)$ is sigmoid to limit the values within the range of $0 \leq G_{\omega,k} \leq 1$ [14]. Since the PSA only manipulates the amplitude spectrum, the PSA cannot perfectly retrieve the target signal when the phase spectrum of $S_{\omega,k}$ does not coincide with that of $N_{\omega,k}$.

On the DFT-domain, we need to manipulate both amplitude and phase of the spectrum to improve the performance of sound source enhancement [20]. To achieve this, a complex-valued T-F mask must be used. The cIRM is a complex-valued T-F mask defined as

$$G^{\mathrm{cIRM}}_{\omega,k} = \frac{S_{\omega,k}}{X_{\omega,k}} = G^{\mathrm{cIRM}}_{\Re,\omega,k} + iG^{\mathrm{cIRM}}_{\Im,\omega,k} \tag{6}$$

where $i^2 = -1$, $\Re$ and $\Im$ denote the real and the imaginary parts of a complex number, respectively. Since typical neural-networks cannot deal with complex values, Williams *et al.* have constructed $\mathcal{M}(\boldsymbol{\phi}_k|\Theta)$ to jointly predict the real and the imaginary parts of the cIRM as [21, 22]

$$(\hat{\boldsymbol{G}}^{\mathrm{cIRM}}_{\Re,k}, \hat{\boldsymbol{G}}^{\mathrm{cIRM}}_{\Im,k})^\top = \mathcal{M}(\boldsymbol{\phi}_k|\Theta), \tag{7}$$

where $\hat{\boldsymbol{G}}^{\mathrm{cIRM}}_{\Re,k} \in \mathbb{R}^\Omega$ and $\hat{\boldsymbol{G}}^{\mathrm{cIRM}}_{\Im,k} \in \mathbb{R}^\Omega$ are vectorized cIRMs of the real and the imaginary parts, respectively[1]. Namely, a complex-valued mask can be estimated by dealing with a complex number as two real numbers. However, although amplitude and phase of the spectrum would not be independent variables [28], the number of output units doubles, *i.e.* 2Ω, and an even more massive amount of training data is required to avoid over-fitting. Therefore, it might be better to use more efficient signal representation than the DFT spectrum for acoustic signal processing using neural networks.

3. PROPOSED METHOD

We propose an end-to-end DNN-based source enhancement with a T-F mask in the MDCT-domain. In Sec. 3.1, the MDCT on which the proposed method is based is briefly introduced in matrix notation. After that, the proposed method is described in Sec. 3.2.

3.1. Modified discrete cosine transform in matrix notation

First, the observed signal x_t is separated into K-blocks of length L without overwlapping and k-th block observation is defined as

$$\mathbf{x}_k := (x_{(k-1)L+1}, x_{(k-1)L+2}, ..., x_{(k-1)L+L})^\top. \tag{8}$$

Then, the k-th time-frame signal is generated by concatenating two blocks, and the MDCT and its inverse-MDCT (IMDCT) of k-th frame are expressed in matrix notation respectively as

$$\mathbf{X}^C_k = \mathbf{C}\mathbf{W}\begin{bmatrix}\mathbf{x}_{k-1}\\ \mathbf{x}_k\end{bmatrix}, \tag{9}$$

$$\mathbf{x}^C_k := \begin{bmatrix}\mathbf{x}^{(C1)}_k\\ \mathbf{x}^{(C2)}_k\end{bmatrix} = \mathbf{W}\mathbf{C}^\top\mathbf{X}^C_k, \tag{10}$$

where $\top$ is transposition, and $\mathbf{X}^C_k := \left(X^C_{1,k}, ..., X^C_{L,k}\right)^\top$ is the vectorized MDCT coefficients. The matrix $\mathbf{C} \in \mathbb{R}^{L\times 2L}$ is the MDCT transformation matrix the (p, q) element of which is defined as

$$C_{p,q} = \sqrt{\frac{2}{L}} \cos\left[\frac{\pi}{L}\left(p + \frac{1}{2}\right)\left(q + \frac{L+1}{2}\right)\right], \tag{11}$$

and the diagonal matrix $\mathbf{W} \in \mathbb{R}^{2L\times 2L}$ corresponds to the analysis/synthesis window commonly defined as

$$W_{\ell,\ell} = \sin\left[\left(\ell + \frac{1}{2}\right)\frac{\pi}{2L}\right]. \tag{12}$$

Since $\mathbf{C}$ is a $L \times 2L$ matrix, the IMDCT vector components $\mathbf{x}^{(C1)}_k$ and $\mathbf{x}^{(C2)}_k$ are corrupted by time-domain aliasing. Fortunately, these aliasings are canceled and the original signal is perfectly reconstructed by adding two subsequent IMDCT vector components as

$$\mathbf{x}_k = \mathbf{x}^{(C2)}_k + \mathbf{x}^{(C1)}_{k+1} = \mathbf{O}\begin{bmatrix}\mathbf{x}^C_k\\ \mathbf{x}^C_{k+1}\end{bmatrix}, \tag{13}$$

which is called time-domain aliasing cancellation (TDAC). Here $\mathbf{O} = [\mathbf{0}, \boldsymbol{I}, \boldsymbol{I}, \mathbf{0}]$ is the overwlap-add (OLA) matrix, and $\mathbf{0}$ and $\boldsymbol{I}$ are the $L \times L$ zero and identity matrices, respectively.

[1] In a practical case, the compressed cIRM using a hyperbolic tangent is estimated to stabilize the training [21]

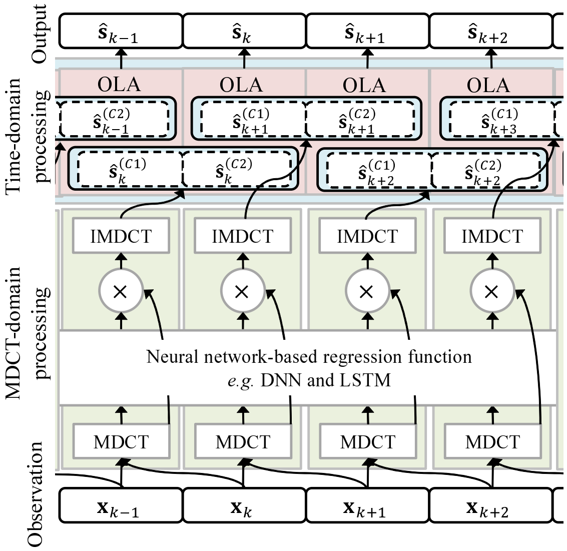

Fig. 1. Proposed end-to-end source enhancement procedure expressed in (18).

3.2. End-to-end DNN-based source enhancement in MDCT-domain

Since MDCT-spectra are real-valued, a real-valued T-F mask in the MDCT-domain enables both amplitude and phase of the spectrum to be manipulated by using a real-valued T-F mask rather than manipulating only the amplitude of the spectrum. In addition, since the size of the MDCT transformation matrix $\mathbf{C}$ is $L \times 2L$, the number of DNN output units can be half the frame size of the MDCT, *i.e.*, the same degree of freedom as the PSA. However, directly manipulating the MDCT spectrum by using a T-F mask causes time-domain aliasing [26]. To reduce the time-domain aliasing, we use an objective function in the time-domain to train DNN by extending DNN-based source enhancement to an end-to-end system as shown in Fig.1.

Here, we define the T-F mask in the MDCT-domain and its processing in the same way as Kuech and Elder [26]

$$G^C_{\ell,k} = \frac{S^C_{\ell,k}}{X^C_{\ell,k}} \tag{14}$$

$$\hat{S}^C_{\ell,k} = G^C_{\ell,k} X^C_{\ell,k}, \tag{15}$$

where $G^C_{\ell,k}$ is a T-F mask in the MDCT-domain and $\hat{S}^C_{\ell,k}$ is the MDCT-spectrum of the output signal. To estimate the T-F mask, we use neural networks $\mathcal{M}(\boldsymbol{\psi}_k|\Theta)$ in the same fashion as (4)

$$\hat{\mathbf{G}}^C_k := (\hat{G}^C_{1,k}, ..., \hat{G}^C_{L,k})^\top = \mathcal{M}(\boldsymbol{\psi}_k|\Theta), \tag{16}$$

where $\boldsymbol{\psi}_k$ is an input acoustic feature of k-th frame in the MDCT domain. Thus, (15) can be rewritten in vectorized notation as

$$\hat{\mathbf{S}}^C_k = (\mathcal{M}(\boldsymbol{\psi}_k|\Theta) + \epsilon) \odot \mathbf{X}^C_k, \tag{17}$$

where $\hat{\mathbf{S}}^C_k := (\hat{S}^C_{1,k}, ..., \hat{S}^C_{L,k})^\top$ and ϵ is a flooring parameter to avoid musical-noise [29]. In fact, the T-F mask in the MDCT-domain defined by (14) takes values in the range $(-\infty, \infty)$, because $S^C_{\ell,k}$ and $X^C_{\ell,k}$ also take values in the range $(-\infty, \infty)$. The large value range may complicate both training DNN and estimating the T-F mask. To limit the values within the range of $0 \leq G^C_{\ell,k} \leq 1$, we use sigmoid as the activation function of output layer of $\mathcal{M}(\boldsymbol{\phi}_k|\Theta)$.

As an objective function for training DNN parameters Θ, use of frame-by-frame MSE between $\hat{\mathbf{S}}^C_k$ and MDCT coefficients of the target signal $\mathbf{S}^C_k$ is an intuitive method like (5). However, it would actually be difficult because the T-F mask in the MDCT-domain corrupts the characteristics of TDAC, *i.e.*, the effect of the time-domain aliasing remains in the output signal [26]. One solution to reduce the time-domain aliasing is to train DNN using an objective function defined in the time-domain. Since the output signal is transformed back to the time-domain by OLA as (13), the error caused by the time-domain aliasing can be mitigated while the estimation error is also minimized. To calculate errors in the time-domain, we build an end-to-end DNN-based source enhancement in the MDCT-domain by using (9),(10),(13) and (17) as

$$\hat{\mathbf{s}}_k = \mathbf{O} \begin{bmatrix} \mathbf{W}\mathbf{C}^\top \left((\mathcal{M}(\boldsymbol{\psi}_k|\Theta) + \epsilon) \odot \mathbf{C}\mathbf{W} \begin{bmatrix} \mathbf{x}_{k-1} \\ \mathbf{x}_k \end{bmatrix} \right) \\ \mathbf{W}\mathbf{C}^\top \left((\mathcal{M}(\boldsymbol{\psi}_{k+1}|\Theta) + \epsilon) \odot \mathbf{C}\mathbf{W} \begin{bmatrix} \mathbf{x}_{k} \\ \mathbf{x}_{k+1} \end{bmatrix} \right) \end{bmatrix}. \tag{18}$$

Then, for an objective function to minimize the time-domain error, we use mean-absolute-error (MAE), which is used as an objective function in a previous end-to-end source enhancement system [30] as

$$\mathcal{J}(\Theta) = \sum_{k=2}^{K-1} ||\mathbf{s}_k - \hat{\mathbf{s}}_k||_1, \tag{19}$$

where $\mathbf{s}_k$ is the k-th block of target source likewise as (8).

4. EXPERIMENTS

We conducted objective experiments to evaluate the performance of the proposed method. For methods to compare with the proposed method, we used the PSA [14] and cIRM [21] for DFT-domain DNN-based source enhancement and the speech enhancement generative adversarial network (SEGAN) [30] for time-domain end-to-end source enhancement.

4.1. Experimental conditions

4.1.1. Dataset

As the training dataset of the target source, the ATR Japanese speech database [31] was used. The dataset consisted of 6640 utterances spoken by 11 male and 11 female speakers. The utterances were randomly classified into two sets: a training set consisting of 5976 speech files and a validation set including 664 speech files, which is used for an early-stopping algorithm described later. As the training dataset of noise, a noise dataset of CHiME-3 was used that consisted of four types of background noise: *cafes*, *street junctions*, *public transport (buses)*, and *pedestrian areas* [32]. The noisy signals for training/validation dataset were formed by mixing clean speech utterances with the noise at signal-to-noise ratio (SNR) levels of -6, 0, 6, and 12 dB. As the test datasets, a Japanese speech database consisting of 300 utterances spoken by 3 males and 3 females was used for target source dataset, and an ambient noise database recorded at airports, amusement parks, offices and party rooms were used as the noisy dataset. All files were recorded at the sampling rate of 16 kHz.

Table 1. Network architectures.

Layer num.	Type	Size, (activation)
		DNN
Input	Fully	$704 \rightarrow 512$, (ReLU)
Hidden 1	Fully	$512 \rightarrow 512$, (ReLU)
Hidden 2	Fully	$512 \rightarrow 512$, (ReLU)
Hidden 3	Fully	$512 \rightarrow 512$, (ReLU)
Hidden 4	Fully	$512 \rightarrow 512$, (ReLU)
Output	Fully	$512 \rightarrow \begin{cases} 64 \text{ (sigmoid)} & \text{(PSA, Prop.)} \\ 64 \times 2 \text{ (linear)} & \text{(cIRM)} \end{cases}$
		LSTM
Input	Fully	$704 \rightarrow 512$, (ReLU)
Hidden 1	LSTM	$512 \rightarrow 512$, (N/A)
Hidden 2	LSTM	$512 \rightarrow 512$, (N/A)
Output	Fully	$512 \rightarrow \begin{cases} 64 \text{ (sigmoid)} & \text{(PSA, Prop.)} \\ 64 \times 2 \text{ (linear)} & \text{(cIRM)} \end{cases}$

4.1.2. *DNN architecture and setup*

The performances of the proposed method and the DFT-domain conventional methods were evaluated on two types of neural networks: a fully connected DNN and an LSTM. The DNN had 4 hidden layers and 512 hidden units, and the LSTM has 2 LSTM-layers and 512 cells. The rectified linear unit (ReLU) was used as the activation functions of hidden layer. To avoid over-fitting, DNN outputs of the proposed and DFT-domain conventional methods were compressed by using a 64-dimensional Mel-transformation matrix, and the estimated T-F masks were transformed into a linear frequency domain by using the Mel-transform's pseudo-inverse [12]. For the proposed method, the block-size was $L = 256$ samples and the frame-size was 512 samples. Input feature of DNN and LSTM were multi-frame log Mel filterbank coefficients of MDCT spectrum defined as

$$\boldsymbol{\psi}_k := \left(\ln\left[\text{Mel}\left[\text{Abs}\left[\mathbf{X}^C_{k-R}\right]\right]\right], ..., \ln\left[\text{Mel}\left[\text{Abs}\left[\mathbf{X}^C_{k+R}\right]\right]\right]\right),$$

where the context window size was $R = 5$, Mel[$\cdot$] and Abs[$\cdot$] denotes the operation of 64-dimensional Mel matrix multiplication and element-wise absolute-value, respectively. A flooring parameter was $\epsilon = 0.1$. For the DFT-domain methods, 64-dimensional multi-frame log Mel filterbank coefficients of DFT spectrum were used as an input feature $\boldsymbol{\phi}_k$. The frame size of the DFT was 512 samples, and the frame was shifted by 256 samples. All the above-mentioned architectures are summarized in Table 1.

The Adam method [33] was used as a gradient method, and the mini-batch size was 50 utterances. To train both methods, DNN and LSTM were trained by layer-by-layer supervised pre-training. All input vectors were mean-and-variance normalized using the training data statistics. An early-stopping algorithm [8] was used with an initial step-size 10^{-4} and step-size threshold 10^{-7}, and L_2 normalization with parameter 10^{-4} was used as a regularization algorithm. For SEGAN, the architecture and training procedure described by Pascual et al. [30] were used.

4.2. Objective evaluations

The source enhancement performance of the proposed method was compared with those of conventional methods using three objective measurements: the SDR, the STOI [27], and the PESQ. Table 2 shows the evaluation results. The PESQ and the STOI of the proposed method were always higher than those of the conventional methods irrespective of the input SNR conditions or selected network. In terms of the SDR, the proposed method outperformed the conventional methods when the input SNR was reasonably high (*i.e.*, > -6 dB). In addition, significant differences were observed for almost scores and input SNR conditions. These results allow us to conclude that the MDCT has a high affinity for DNN-based source enhancement.

Table 2. Experimental results. Asterisks indicate scores that were significantly higher than scores provided by the second-placed method using the same network in a paired one-sided t-test ($\alpha = 0.05$).

Input SNR	Network	T-F mask	SDR	STOI	PESQ
-6 dB	SEGAN	-	1.19	64.7	1.26
	DNN	PSA	5.57	75.1	1.87
		cIRM	4.58	75.6	1.77
		Proposed	***5.97**	***76.5**	***1.94**
	LSTM	PSA	***6.73**	78.7	2.02
		cIRM	5.35	77.9	1.95
		Proposed	6.43	***79.6**	**2.03**
0 dB	SEGAN	-	8.40	83.3	1.95
	DNN	PSA	10.61	85.9	2.38
		cIRM	9.84	86.1	2.28
		Proposed	***11.70**	***89.0**	***2.50**
	LSTM	PSA	11.86	89.5	2.54
		cIRM	10.55	88.3	2.46
		Proposed	***12.09**	***90.6**	**2.57**
6 dB	SEGAN	-	14.06	92.2	2.39
	DNN	PSA	15.02	92.3	2.76
		cIRM	13.58	92.2	2.72
		Proposed	***16.63**	***94.8**	***2.92**
	LSTM	PSA	16.40	94.8	2.92
		cIRM	14.56	93.8	2.87
		Proposed	***16.97**	***95.5**	***2.97**
12 dB	SEGAN	-	18.73	95.7	2.72
	DNN	PSA	18.88	95.9	3.09
		cIRM	16.00	95.3	3.12
		Proposed	***21.07**	***97.3**	***3.30**
	LSTM	PSA	20.60	97.2	3.25
		cIRM	17.43	96.4	3.22
		Proposed	***21.50**	***97.7**	***3.34**

5. CONCLUSIONS

In this study, we proposed an end-to-end DNN-based source enhancement method on the basis of T-F mask processing in the MDCT-domain. The key idea of this study was the use of MDCT as a frequency transformation instead of the DFT, which enables us to i) manipulate both amplitude and phase of the spectrum by using a real-valued T-F mask and ii) achieve end-to-end training using DNN output units numbering the same as or fewer than those of previous DNN-based source enhancement algorithms. Since T-F mask processing in the MDCT domain causes time-domain aliasing [26], we defined an objective function in the time-domain to reduce the time-domain aliasing for optimizing the networks. Experimental results showed that the proposed method significantly outperformed the conventional methods in terms of the SDR, STOI and PESQ scores in almost all SNR conditions. Thus, we conclude that the MDCT has a high affinity for DNN-based source enhancement.

6. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Eds., "Speech enhancement," Springer, 2005.

[2] T. Yoshioka, N. Ito, M. Delcreix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Dabian, M. Espi, T. Higuchi, A. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,," in *Proc. ASRU*, 2015.

[3] T. Ochiai, S. Watanabe, T. Hori, and J. Hershey, "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017.

[4] K. Kobayashi, Y. Haneda, K. Furuya, and A. Kataoka, "A hands-free unit with noise reduction by using adaptive beamformer," *IEEE Trans. on Consumer Electronics*, pp.116–122, 2008.

[5] B. C. J. Moore, "Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms," *Speech Communication*, pp.81–91, 2003.

[6] D. L. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, pp. 332–353, 2008.

[7] T. Zhang, F. Mustiere, and C. Micheyl, "Intelligent hearing aids: The next revolution," in *Proc. EMBC,* 2016.

[8] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi and H. Ohmuro, "Informative acoustic feature selection to maximize mutual information for collecting target sources," *IEEE/ACM Trans. Audio, Speech and Language Processing*, pp.768–779, 2017.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech and Language Processing*, pp.1109–1121, 1984.

[10] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. Audio, Speech and Language Processing*, pp.1240–1250, 2013. *Speech communication*, pp.229–244, 2012.

[11] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. WASPAA,* 2003.

[12] F. Weninger, J. R. Hershey, J. L. Roux and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, 2014.

[13] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. ICASSP*, 2015.

[14] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015.

[15] 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*, pp.7–19, 2015.

[16] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi and Y. Hioka, "Pinpoint extraction of distant sound source based on DNN mapping from multiple beamforming outputs to prior SNR" in *Proc. ICASSP*, 2016.

[17] J. Hershy, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016.

[18] Q. Liu, W. Wang, P. J. B Jackson, and Y. Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions," in *Proc. EUSIPCO*, 2017.

[19] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi and Y. Haneda, "DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *Proc. ICASSP*, 2017.

[20] K. Paliwal, K. Wojcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, pp. 465–494, 2010.

[21] D. S. Williamson, Y. Wang and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* pp.483–492, 2016.

[22] D. S. Williamson and D. L. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* pp.1492–1501, 2017.

[23] E. Variani, T. N. Sainath, I. Shafran, and M. Bacchiani, "Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling," in *Proc INTERSPEECH*, 2016.

[24] J. P. Prince and A. B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE/ACM Transactions on Audio, Speech, and Signal Processing,* pp.1153–1161, 1986.

[25] Y. Wang and M. Vilermo, "Modified discrete cosine transform—Its implications for audio coding and error concealment," *J. Audio Eng. Soc.,* pp.52–61, 2003.

[26] F. Keuch and B. Elder, "Aliasing reduction for modified discrete cosine transform domain filtering and its application to speech enhancement," In *Proc WASPAA*, 2007.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, pp.2125–2136, 2011.

[28] S. Shimauchi, S. Kudo, Y. Koizumi, and K. Furuya, "On relationships between amplitude and phase of short-time Fourier transform," in *Proc. ICASSP*, 2017.

[29] L. Lightburn, E. D. Sena, A. Moore, P. A. Naylor, M. Brookes, "Improving the perceptual quality of ideal binary masked speech," in *Proc. ICASSP*, 2017.

[30] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," In *Proc INTERSPEECH*, 2017.

[31] 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*, pp.357–363, 1990.

[32] J. Barker, R. Marxer, E. Vincent and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baseline," in *Proc. ASRU*, 2015.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc ICLR*, 2015.