DNN-BASED SOURCE ENHANCEMENT SELF-OPTIMIZED BY REINFORCEMENT LEARNING USING SOUND QUALITY MEASUREMENTS

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

[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

We investigated whether a deep neural network (DNN)-based source enhancement function can be self-optimized by reinforcement learning (RL). The use of a DNN is a powerful approach to describing the relationship between two sets of variables and can be useful for source enhancement function design. By training the DNN using a huge amount of training data, sound quality of output signals are improved. However, collecting a huge amount of training data is often difficult in practice. To use limited training data efficiently, we focus on the "self-optimization" of DNN-based source enhancement function in which RL is commonly utilized in the development of game playing computers. As a reward for RL, quantitative metrics that reflect a human's perceptual score (*perceptual score*), e.g., perceptual evaluation methods for audio source separation (PEASS), are utilized. To investigate whether the sound quality is improved by RL-based source enhancement, subjective tests were conducted. It was confirmed that the output sound quality of the RL-based source enhancement function improved as the number of iterations was increased and finally outperformed the conventional method.

Index Terms— Sound source enhancement, Time-frequency mask, Reinforcement learning, Sound quality and perceptual score.

1. INTRODUCTION

Sound source enhancement has been studied for many years [1, 2, 3, 4, 5] because of high demand for its use for various practical applications such as automatic speech recognition [6, 7] , hearing aids, and immersive audio field representation [8, 9]. The goal of this study is to collect target sound sources in noisy environments. To achieve this goal, time-frequency (T-F) masking has been commonly employed, e.g., Wiener filtering [2]. To accurately estimate T-F masks, various approaches have been developed including multi-channel approaches, e.g., [3, 5], and statistical approaches, e.g., [4, 9].

Recently, deep neural network (DNN)-based sound source enhancement has been actively studied [7, 10, 11, 12, 13, 14, 15, 16]. In our previous works [12, 13], DNNs were utilized as a mapping function to estimate T-F masks. Hershey *et al.* utilized a DNN as a clustering function to estimate ideal binary masks [14]. To improve the output signal quality of DNN-based source enhancement, a huge amount of training data, which is composed of, e.g., observed signals and supervised T-F masks, is needed. However, collecting a huge amount of training data is often difficult in practice, which may hinder the improvement of output signal quality.

To use limited training data efficiently, our strategy is to "self-optimize" the source enhancement function—no explicit supervised T-F mask is provided. As a self-optimization approach, reinforcement learning (RL) is commonly employed, especially in the development of game playing computers [17, 18, 19]. In RL, instead of defining supervised output, a *reward* needs to be defined; the reward indicates the validity of adopted tactics [20]. In the development of game playing computers, a game playing function, i.e., playing policy, is successfully self-optimized by designing the reward from an explicit scoring function, e.g. win/lose or game score.

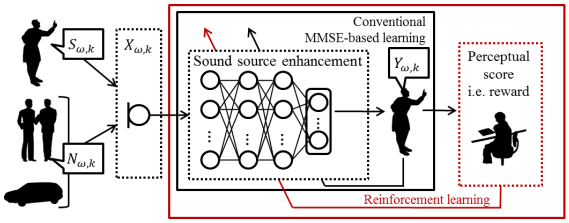

Fig. 1. Concept of RL-based source enhancement

If an appropriate reward could be designed similar to the case of game playing computers, a DNN-based source enhancement function may be able to be self-optimized by RL. In this study, RL-based self-optimization is applied to the training of a DNN-based source enhancement function by using a quantitative metric that reflects a human's perceptual score (*perceptual score*) as the reward, instead of explicitly giving supervised T-F masks (Fig. 1). In this paper, we focus on investigating if a DNN-based source enhancement function can be self-optimized by RL with a reward calculated from a conventional perceptual score, e.g., the perceptual evaluation of speech quality (PESQ) [21] and perceptual evaluation methods for audio source separation (PEASS) [22]. To investigate the validity of the proposed RL-based source enhancement, output sound quality was evaluated by both objective and subjective evaluations.

The rest of this paper is organized as follows. Section 2 introduces the conventional DNN-based sound-source enhancement. Then, in Section 3, a framework of the RL-based self-optimization of source enhancement function is proposed. After investigating the sound quality of output signals through several objective and subjective tests in Section 4, we conclude this paper with some remarks in Section 5.

2. CONVENTIONAL METHOD

2.1. Sound source enhancement using time-frequency masking

In this paper, we consider the problem of determining a target source $S_{\omega,k}$ surrounded by ambient noise $N_{\omega,k}$. A signal observed with a

single microphone $X_{\omega,k}$ is expressed as

$$X_{\omega,k} = H_\omega S_{\omega,k} + N_{\omega,k}, \tag{1}$$

where $\omega = \{1, 2, ..., \Omega\}$ and $k = \{1, 2, ..., K\}$ denote the frequency and time indices, respectively. H_ω is the transfer function from the target source to the microphone.

In sound source enhancement using T-F masking, the output signal $Y_{\omega,k}$ is obtained by multiplying a T-F mask to $X_{\omega,k}$ as

$$Y_{\omega,k} = G_{\omega,k} X_{\omega,k}, \tag{2}$$

where $G_{\omega,k}$ is the T-F mask such as a frame-wise Wiener filter [2] and ideal ratio mask (IRM) [7]. The ideal frame-wise Wiener filter $G^{\text{ideal}}_{\omega,k}$ [2] can be calculated by

$$G^{\text{ideal}}_{\omega,k} = \frac{|H_\omega S_{\omega,k}|^2}{|H_\omega S_{\omega,k}|^2 + |N_{\omega,k}|^2} \tag{3}$$

by assuming that the target source and surrounding noise are mutually uncorrelated. However, $H_\omega S_{\omega,k}$ and $N_{\omega,k}$ in (3) would be unknown in practice. Thus, we need to estimate $H_\omega S_{\omega,k}$ and $N_{\omega,k}$ from $X_{\omega,k}$ to obtain $G_{\omega,k}$ and $Y_{\omega,k}$.

2.2. Time-frequency mask estimation through DNN-mapping

The general statistical source enhancement approach estimates vectorized T-F masks for all frequency bins [1] $\boldsymbol{G}_k = (G_{1,k}, ..., G_{\Omega,k})^\top$, here, $\top$ denotes transposition. In DNN-mapping-based source enhancement (named *DNN-mapping* hereafter) [7, 10, 11, 12, 13], $\boldsymbol{G}_k$ is estimated with DNN parameter $\Theta = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}, l \in (2, ..., L)\}$ as

$$\hat{\boldsymbol{G}}_k \leftarrow \mathbf{W}^{(L)} \boldsymbol{z}_k^{(L-1)} + \mathbf{b}^{(L)}, \tag{4}$$

$$\boldsymbol{z}_k^{(l)} = \sigma_\theta \left\{ \mathbf{W}^{(l)} \boldsymbol{z}_k^{(l-1)} + \mathbf{b}^{(l)} \right\}, \tag{5}$$

where L, $\mathbf{W}^{(l)}$, and $\mathbf{b}^{(l)}$ are the number of layers, the weight matrix, and the bias vector, respectively. The function σ_θ is a nonlinear activation function, such as a sigmoid function. The input vector $\boldsymbol{x}_k$, which is passed to the first layer of the network as $\boldsymbol{z}_k^{(1)} = \boldsymbol{x}_k$, is obtained by concatenating several frames of observation features to account for previous and future frames, as

$$\boldsymbol{x}_k = (\boldsymbol{X}_{k-P}, ..., \boldsymbol{X}_k, ..., \boldsymbol{X}_{k+P})^\top, \tag{6}$$

$$\boldsymbol{X}_k = (X_{1,k}, ..., X_{\Omega,k}), \tag{7}$$

where P is the context window size.

The MMSE-based objective function is widely used to train Θ, namely the mean square error between DNN output $\hat{G}_{\omega,k}$ and the ideal T-F mask $G^{\text{ideal}}_{\omega,k}$ is minimized. The objective function is designed by assuming that the target source is collected clearly by maximizing the SNR of the output signal given by

$$\Theta \leftarrow \arg\min_{\Theta} \frac{1}{K} \sum_{k=1}^{K} \sum_{\omega=1}^{\Omega} \left| G^{\text{ideal}}_{\omega,k} - \hat{G}_{\omega,k} \right|^2, \tag{8}$$

which is solved by using the back-propagation algorithm [23].

Since the number of parameters in Θ is large, we need to collect a huge dataset composed of $\boldsymbol{x}_k$ and $G^{\text{ideal}}_{\omega,k}$ to improve the output signal quality of DNN-mapping. However, collecting a huge amount of training data is often difficult in practice, which may hinder the improvement of the output signal quality.

[1] Instead of directly estimating the T-F mask, some approaches estimate log-power spectra of a target source [10] or log SNR [12].

3. PROPOSED METHOD

In order to use limited training data efficiently, we apply RL-based self-optimization to design a source enhancement function as shown in Fig. 1. In this paper, a quantitative metrics that reflect a human's perceptual score such as the PESQ [21] and PEASS [22] are utilized as the *reward* [2] of the RL, and the DNN-based source enhancement function is sequentially optimized so as to maximize the reward.

3.1. Frame work of reinforcement learning for source enhancement function optimization

Figure 2 shows the overall procedure of the proposed speech enhancement method. Generally, the RL schema requires a finite number of *actions* $\mathcal{A} = \{a_1, ..., a_A\}$ to be predefined [20]. In our problem setting, an action a is defined by the a-th T-F mask $\boldsymbol{\mathcal{G}}_a = (G_{1,a}, ..., G_{\Omega,a})^\top$; thus, a finite number of actions can be given by T-F mask templates $\boldsymbol{\mathcal{G}}_{1,...,A}$. The action is selected in accordance with the observation $\boldsymbol{x}_k$ and its own selection policy $\mathcal{Q}(\boldsymbol{x}_k, a)$. To select a suitable action, $\mathcal{Q}(\boldsymbol{x}_k, a)$ should be appropriately designed. Namely, optimized $\mathcal{Q}(\boldsymbol{x}_k, a)$ would take a high probability value when the T-F mask $\boldsymbol{\mathcal{G}}_a$ leads to an output signal with high sound quality.

$$\hat{\boldsymbol{G}}_k \leftarrow \boldsymbol{\mathcal{G}}_{a_k}, \tag{9}$$

$$a_k \leftarrow \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\boldsymbol{x}_k, a). \tag{10}$$

From (9)(10), it can be regarded that $\mathcal{Q}(\boldsymbol{x}_k, a)$ discriminates the optimal template $\boldsymbol{\mathcal{G}}_{a_k}$. To accurately select the optimal template, $\mathcal{Q}(\boldsymbol{x}_k, a)$ is implemented by DNN whose non-linear function of the output layer is a softmax function as

$$\mathcal{Q}(\boldsymbol{x}_k, a) = \frac{\exp(z^{(L)}_{k,a})}{\sum_{i=1}^{A} \exp(z^{(L)}_{k,i})}, \tag{11}$$

$$\boldsymbol{z}_k^{(L)} = \mathbf{W}_q^{(L)} \boldsymbol{z}_k^{(L-1)} + \mathbf{b}_q^{(L)} \tag{12}$$

because DNN is one of the most powerful discriminant functions. Here, $\boldsymbol{z}^{(L)} = (z^{(L)}_{k,1}, ..., z^{(L)}_{k,A})^\top$. Hereafter, sound source enhancement using T-F mask estimated by (9)-(12) is named as *DNN-RL*.

In the RL schema, the parameter of DNN-RL $\Theta_q = \{\mathbf{W}_q^{(l)}, \mathbf{b}_q^{(l)}\}$ is optimized by maximizing the "reward", also known as Q-learning. Thus, in order to design Θ_q appropriately, it is necessary to design a reward that accurately evaluates the sound quality of the output signal processed by the selected T-F masks.

3.2. Reward design for Q-learning

Although using a perceptual score $\mathcal{Z}$ directly as the reward for Q-learning would be an intuitive way, it would be in fact difficult because $\mathcal{Z}$ is affected not only by the performance of the source enhancement but also by the noise environment, such as an environment with a high or low SNR. To avoid being affected by such external factors, the relative value between the perceptual score calculated from the output of the DNN-RL $\mathcal{Z}$ and that of the DNN-mapping $\mathcal{Z}^{\text{DNN}}$ is calculated as

$$\mathcal{R} = \tanh\left\{ \alpha \left(\mathcal{Z} - \mathcal{Z}^{\text{DNN}} \right) \right\}, \tag{13}$$

[2] The reward indicates the validity of adopted tactics [20]. In our problem setting, the reward specifies the output signal quality of the RL-based source enhancement function.

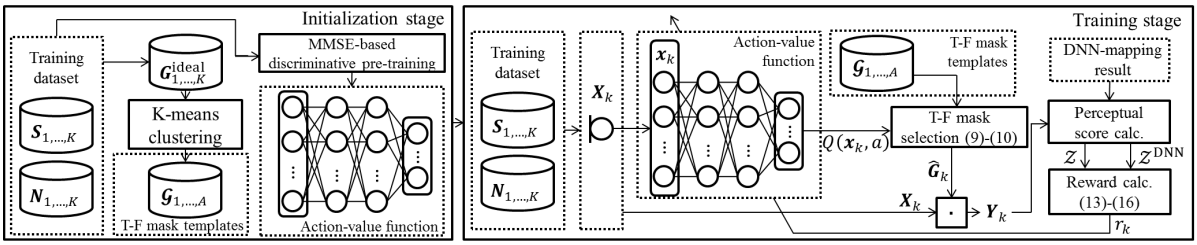

Fig. 2. Overview of procedures in proposed method

where $\alpha > 0$ is a scaling parameter of $\mathcal{Z}$. This relative value is inspired by the victory or defeat in game playing [19]. If $\mathcal{Z}$ is higher than $\mathcal{Z}^{\text{DNN}}$, i.e., DNN-RL won against the DNN-mapping, the reward takes a positive value. If $\mathcal{Z}$ is lower than $\mathcal{Z}^{\text{DNN}}$, i.e., DNN-RL lost to the DNN-mapping, the reward takes a negative value, i.e., a penalty. Hyperbolic tangent clipping limits the scale of the perceptual score and aims to avoid a large gradient value described later.

In addition, the reward should be varied with time k because a T-F mask is also time-variant. However, most existing perceptual scores, e.g., PESQ, cannot be calculated for each time k because their calculation requires multiple frames. Hence, a local *mis-action*, i.e., fallacious T-F mask selection, would also be given a low-perceptual score. To design a time varying reward, we utilize a time-weight E_k in the reward calculation as

$$r_k = \begin{cases} (1 - E_k)\mathcal{R} & (\mathcal{R} > 0) \\ E_k \mathcal{R} & (\text{other}) \end{cases}, \tag{14}$$

$$E_k = \frac{\tilde{E}_k}{\max_{k \in K}(\tilde{E}_k)}, \tag{15}$$

$$\tilde{E}_k = \sum_{\omega=1}^{\Omega} \left| \ln |Y_{\omega,k}| - \ln |H_\omega S_{\omega,k}| \right|^2. \tag{16}$$

As can be seen in (14)–(16), the time-weight $0 < E_k < 1$ is the normalized squared error between output $Y_{\omega,k}$ and target $S_{\omega,k}$. The square error $\tilde{E}_k$ around local mis-actions would become sufficiently higher than that of other times. Thus, the normalized square error in (14) works as a penalty for local mis-actions.

By using r_k, the target value of action-value function $\tilde{\mathcal{Q}}(\boldsymbol{x}_k, a_k)$ is calculated as

$$\tilde{\mathcal{Q}}(\boldsymbol{x}_k, a_k) = \begin{cases} r_k + \max_{a \in \mathcal{A}} \mathcal{Q}(\boldsymbol{x}_k, a) & (\mathcal{R} > 0) \\ \mathcal{Q}(\boldsymbol{x}_k, a_k) & (\text{other}) \end{cases}, \tag{17}$$

where if $a_k \neq a_k^{\text{MMSE}}$, $\tilde{\mathcal{Q}}(\boldsymbol{x}_k, a_k^{\text{MMSE}})$ is calculated by

$$\tilde{\mathcal{Q}}(\boldsymbol{x}_k, a_k^{\text{MMSE}}) = \begin{cases} \mathcal{Q}(\boldsymbol{x}_k, a_k^{\text{MMSE}}) & (\mathcal{R} > 0) \\ \mathcal{Q}(\boldsymbol{x}_k, a_k^{\text{MMSE}}) - r_k & (\text{other}) \end{cases}, \tag{18}$$

where a_k^{MMSE} is the MMSE-sense T-F mask label calculated as

$$a_k^{\text{MMSE}} \leftarrow \arg\min_{a \in \mathcal{A}} \sum_{\omega=1}^{\Omega} \left| |H_\omega S_{\omega,k}| - |\mathcal{G}_{\omega,a} X_{\omega,k}| \right|^2. \tag{19}$$

Since $\mathcal{Q}(\boldsymbol{x}_k, a)$ is an output of the softmax function (11), $\tilde{\mathcal{Q}}(\boldsymbol{x}_k, a)$ is normalized and floored to satisfy $\sum_{i=1}^{A} \tilde{\mathcal{Q}}(\boldsymbol{x}_k, i) = 1$ and $\tilde{\mathcal{Q}}(\boldsymbol{x}_k, a) \geq 0$. The definition of the reward in (18) is not the same as the rewards commonly used in RL schema. The intention behind this definition is to prioritise the DNN-mapping if the RL-based source enhancement fails to outperform the DNN-mapping, i.e., $\mathcal{R} < 0$, which has already been trained with the MMSE-based objective function (8) since it implies that the MMSE-based action a_k^{MMSE} is better than current action a_k. Thus, by subtracting the negative reward r_k from $\mathcal{Q}(\boldsymbol{x}_k, a_k^{\text{MMSE}})$, the action-value function of MMSE-based action a_k^{MMSE} is increased.

3.3. Training procedure

As shown in Fig. 2, the training of our RL-based source enhancement is multistage processing (initialization and training). Here, we describe the details of each processing stage.

In the initialization stage, T-F mask templates $\boldsymbol{\mathcal{G}}_{1,\ldots,A}$ are calculated, and the action-value function is pre-trained. As T-F mask templates $\boldsymbol{\mathcal{G}}_{1,\ldots,A}$, cluster centers calculated by using the k-means algorithm of $\boldsymbol{G}_k^{\text{ideal}}$ are utilized. The action-value function, namely, the DNN parameter Θ_q, is pre-trained in the MMSE sense. In particular, discriminative pre-training [24] to maximize the identification rate of the MMSE-sense T-F mask label a_k^{MMSE} is utilized to initialize Θ_q. Here, the L-th layer parameters $\mathbf{W}^{(L)}, \mathbf{b}^{(L)}$ are initialized with values that follow a normal distribution.

In the training stage, Θ_q is trained to maximize the reward. First, an observation is simulated using a randomly selected target source file and same frame-size of noise source from the training dataset. Next, to accelerate the convergence of DNN training, an output signal is obtained using the T-F mask selected by using the ϵ-greedy strategy [18]; the best action defined by (9)(10) is selected with probability $1 - \epsilon$, and that with probability ϵ a random selection is made instead. Then, the perceptual score and rewards are calculated by (14), and finally, Θ_q is updated to minimize the following criteria.

$$\Theta_q \leftarrow \arg\min_{\Theta} \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{A} \left| \tilde{\mathcal{Q}}(\boldsymbol{x}_k, i) - \mathcal{Q}(\boldsymbol{x}_k, i) \right|^2. \tag{20}$$

In this paper, to minimize (20), the RMSProp algorithm with standard mini-batch stochastic gradient descent (SGD) was used [25].

4. EXPERIMENTS

4.1. Experimental conditions

We conducted objective and subjective evaluations to explore whether the DNN-based source enhancement function can be self-optimized by RL with PESQ or PEASS. As a comparison method, we applied DNN-mapping and an ideal Wiener filter (3) to exhibit the upper limit of performance due to T-F masking.

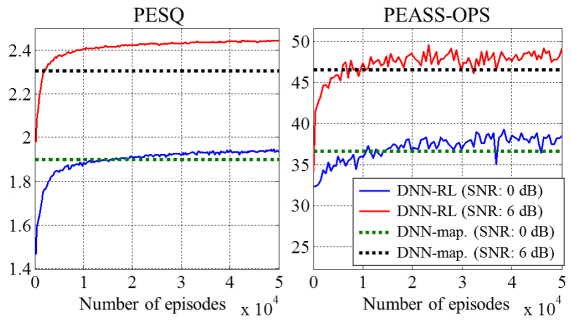

Fig. 3. Perceptual score depending on the number of episodes. The x-axis shows the number of episodes, and the y-axis shows the perceptual score. The solid lines and the dashed lines are DNN-RL and DNN-mapping, respectively.

The training/test dataset of target source and noise were created from the ATR Japanese speech database [26] and the noise dataset of CHiME-3 which consisted of four types of background noise files including cafes, street junctions, public transport (buses), and pedestrian areas [27], respectively. The training dataset consisted of 3316 utterances spoken by 11 males and 11 females at SNRs of 0, 3, and 6 dB. The test dataset consisted of 100 utterances, different speakers and sentences from the training dataset. For training and test, the first half and the last half of the noise files were used, respectively.

For DNN-RL, action-value function had two hidden layers each of which had 64 units. The number of T-F mask templates A was 32. The reward coefficient α for PESQ and PEASS-OPS was 20.0 and 1.0, respectively. The ϵ-greedy parameter ϵ was 0.01. For DNN-mapping, the number of hidden layers was 2 and the number of units in each hidden layer was 128. Instead of direct T-F mask estimation, log-amplitude-spectrum $\ln|S_{\omega,k}|$ was estimated [10]. The dropout algorithm was used as regularization algorithms to avoid over-fitting [10]. The DNN was initialized by discriminative pre-training [24], and a standard mini-batch SGD with momentum was used for fine-tuning. For each method, the context window size $P = 5$. The activation functions of the hidden layers were sigmoid. To avoid over-fitting, $\boldsymbol{X}_k$ and $\boldsymbol{G}_k$ were compressed by $\mathsf{B} = 64$ mel-filterbanks, and the estimated T-F masks were transformed to a linear frequency , i.e., short-time Fourier transform (STFT), domain by spline interpolation. The frame size of the STFT was 512 and the frame was shifted by 256 samples.

4.2. Verification experiments on reinforcement learning

We investigated the relationship between the number of episodes and the perceptual score. An *episode* is a set of training procedure for an utterance; enhancing speech, calculating a perceptual score, and updating the DNN-RL parameters. If $\mathcal{Q}(\boldsymbol{x}_k, a)$ were successfully trained by RL, the perceptual score of the output signal would increase depending on the number of episodes. In this experiment, the noises were mixed to the test dataset at SNRs of 0 and 6 dB.

Figure 3 shows the perceptual score depending on the number of episodes. Both perceptual scores were increased as the number of episodes increased. In addition, since the proposed procedure was specialized to maximize perceptual scores, DNN-RL outperformed DNN-mapping. These results suggest that the proposed procedure is effective at maximizing arbitrary objective measure, such as the perceptual score.

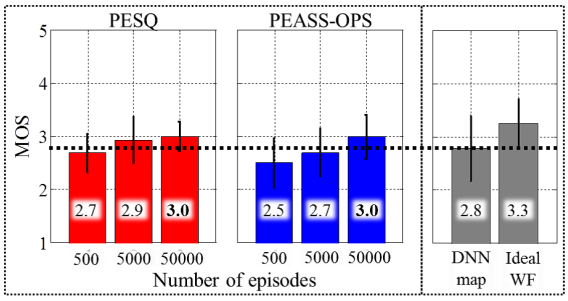

Fig. 4. Results of subjective evaluation. The error bars denote standard deviation. The dashed line shows the MOS of DNN-mapping.

4.3. Subjective evaluation

We conducted a mean-opinion-score (MOS) test to investigate the sound quality of the output signals provided by DNN-RL. Seven participants evaluated sound quality of output signals. We asked the participants to rate the sound using a 5-point scale: 1 – Bad, 3 – Fair, and 5 – Excellent. To remove outliers, top and bottom 5% scores were removed for MOS calculation.

The participants evaluated the ideal Wiener filter, a DNN-mapping, and six DNN-RLs. DNN-RLs consisted of two types of perceptual scores, i.e., PESQ and PEASS-OPS, and three types of episode numbers, i.e., 500, 5,000, and 50,000 episodes. The participants evaluated ten files for each method. In this experiment, the street junction noise was mixed to the test dataset at SNR of 3 dB.

Figure 4 shows the results of the subjective test. The MOSs of DNN-RL were improved according to the number of episodes, and statistically significant differences between 500 episodes and 50,000 episodes were observed in an unpaired one-sided t-test (p-value $= 0.05$). In addition, the MOSs of 50,000 episodes outperformed DNN-mapping and statistically significant differences were observed in an unpaired one-sided t-test (p-value $= 0.05$).

From these results, we found that the DNN-based source enhancement function can be optimized by RL and sound quality of output signals were improved by using a perceptual score as the reward. Thus, it can be concluded that the proposed method can use limited training data efficiently than the conventional training procedure of DNN-based source enhancement function.

5. CONCLUSION

In this paper, we investigated whether a DNN-based source enhancement function could be self-optimized by RL with a reward calculated from a conventional perceptual score, e.g., PESQ and PEASS. To investigate the validity of the proposed method, the output sound quality was evaluated by both objective and subjective evaluations. In these experiments, we found that the DNN-based source enhancement function can be optimized by RL and sound quality of output signals were improved. Thus, it can be concluded that the proposed method can use limited training data efficiently for DNN-based source enhancement function training.

A future prospect of this study is development of new perceptual score which can evaluate sound quality without target source. If such a perceptual score could be developed, training data collection process would become easy because training data of target source would no longer be needed and the range of application of DNN-based source enhancement would be more extended.

6. REFERENCES

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

[2] 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.

[3] R. Zelinski "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP*, pp. 2578 –2581, 1988.

[4] M. Fujimoto, S. Watanabe and T. Nakatani, "Frame-wise model re-estimation method based on Gaussian pruning with weight normalization for noise robust voice activity detection," *Speech communication*, 2012.

[5] 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.

[6] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, pp. 114–126, 2012.

[7] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013.

[8] R. Oldfield, B. Shirley and J. Spille, "Object-based audio for interactive football broadcast," *Multimedia Tools and Applications*, Vol. 74, pp.2717–2741, 2015.

[9] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi and H. Ohmuro, "Integrated approach of feature extraction and sound source enhancement based on maximization of mutual information," in *Proc. ICASSP*, pp. 186–190, 2016.

[10] 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.

[11] D. Bagchi, M. Mandel, Z. Wang, Y. He, A. Plummer and E. F. Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *Proc. ASRU*, 2015.

[12] 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*, pp. 435–439, 2016.

[13] T. Kawase, K. Niwa, K. Kobayashi, and Y. Hioka, "Application of neural network to source PSD estimation for Wiener filter based sound source separation," in *Proc. IWAENC*, 2016.

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

[15] F. Weninger , H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, " Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," in *Proc. LVA/ICA*, 2015.

[16] 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.

[17] G. J. Tesauro, "Temporal difference learning and TD-Gammon," *Communications of the ACM*, pp.58–68, 1995.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou,H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, 518, pp. 529–533, 2015.

[19] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis, " Mastering the game of Go with deep neural networks and tree search," *Nature*, pp.484—489, 2016.

[20] M. Sugiyama, "Statistical reinforcement learning: modern machine learning approaches," *Chapman and Hall/CRC*, 2015.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality Assessment of Narrow-band Telephone Networks and Speech Codecs," in *Proc. ICASSP*, 2001.

[22] V. Emiya, E. Vincent, N. Harlander and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, pp. 2046–2057, 2011.

[23] D.E.Rumelhart, G.E.Hinton, E.Geoffrey and R.J.Williams, "Learning representations by back-propagating errors," *Nature 323* pp.533–536, 1986.

[24] F. Seide, G. Li, X. Chen and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, pp. 24–29, 2011.

[25] T. Tieleman and G. Hinton, "Lecture 6.5 - rmsprop,?f?f *COURSERA: Neural Networks for Machine Learning*, 2012.

[26] 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.

[27] 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