Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
() : –

doi:10.3906/elk-2011-144

Deep Q-network-based noise suppression for robust speech recognition

Tae-Jun PARK, Joon-Hyuk CHANG*
Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea

Received: 27.11.2020 • **Accepted/Published Online:** 02.04.2021 • **Final Version:** ..2021

Abstract: This study develops the deep Q-network (DQN)-based noise suppression for robust speech recognition purposes under ambient noise. We thus design a reinforcement algorithm that combines DQN training with a deep neural networks (DNN) to let reinforcement learning (RL) work for complex and high dimensional environments like speech recognition. For this, we elaborate on the DQN training to choose the best action that is the quantized noise suppression gain by the observation of noisy speech signal with the rewards of DQN including both the word error rate (WER) and objective speech quality measure. Experiments demonstrate that the proposed algorithm improves speech recognition in various noisy conditions while reducing the computational burden compared to the DNN-based noise suppression method.

Key words: Deep Q-network, reinforcement learning, speech recognition, noise suppression, speech enhancement, deep neural network

1. Introduction

Recent years have witnessed major advances of deep neural networks (DNN)-based speech enhancement since the introduction of the regression approach and ideal ratio mask (IRM) [1, 2]. These algorithms successfully suppress the background noise, which is an inevitable problem in real world scenarios. Although speech enhancement techniques are found to be very helpful in tasks such as speech recognition, recognition performance is severely degraded in noisy conditions [3, 4]. It is hence crucial that the speech recognition exhibits a robust performance across ambient noises and benefit many speech technology applications. In [4], the supervised learning approach is proposed to map a set of harmonic features into a pitch based grouping cue for time-frequency (T-F) unit. T-F unit means another domain of a signal converted through the short time Fourier transform from a time domain. The mapping results in an ideal binary mask (IBM) that is used to preserve the speech dominant one in the T-F unit, in which, a value of 1 in the mask indicates that the speech is stronger than noise and a value of 0 otherwise. It is recently found that performing T-F masking in the complex domain is very beneficial to jointly enhance the magnitude and phase response of noisy speech by estimating the complex IRM in the real and imaginary domains [5]. A new speech enhancement algorithm was introduced in [6], this method is based on the adaptive threshold of intrinsic mode functions (IMFs) of noisy signal frames extracted by empirical mode decomposition. Adaptive threshold values are estimated by using the gamma statistical model of Teager energy operated IMFs of noisy speech and estimated noise based on symmetric Kullback–Leibler divergence. Enhanced speech is then obtained from noisy speech by a semisoft thresholding function.

*Correspondence: jchang@hanyang.ac.kr

On the other hand, a regression approach to DNN-based speech enhancement has been proposed to train deep and wide neural networks based on a large training data that encompasses many possible combinations of speech and noise types, which maps directly from observed noisy speech to desired clean speech [7]. It turns out that the annoying musical noise, which is a noise residue that remains after applying speech enhancement, is greatly reduced and thus the enhanced speech shows a higher speech quality in terms of the objective quality metric and listening test. This idea was further extended to cope with adverse conditions and non-stationary noises in real-world scenarios. Recently, a novel structure of network called the SpinalNet [8] was developed, and it was reported to achieve the state-of-the-art (SOTA) performance in the regression task on several well-known benchmark datasets. By observing the recent success of convolutional neural networks (CNN) and the wonderful architecture of human spinal cord Kabir et al., developed a neural network with gradual inputs, named SpinalNet as shown in Figure 1. Among two predominant algorithms [7, 8] we introduced earlier, the SpinalNet has been considered to employ the reinforcement learning (RL) [9] since a huge amount of the training data is necessary to ensure the output speech quality of the DNN-based regression method. In order to efficiently use the training data that is relatively insufficient in absolute amount, RL can be adopted as a promising algorithm since it is self-optimized with a reward which is calculated from the perceptual score in such a way that the amount of computation is greatly reduced [10]. In [11], the Q-learning is proposed to incorporate the feedback from perceptual quality of enhanced speech and then the best T-F mask is selected among the clustered T-F mask templates, supported by the k-means algorithm. Recently, in [12, 13], the RL was further applied to optimize the speech enhancement model for the recognition results. However, since these studies were not designed directly based on the recognition results of automatic speech recognition (ASR), but only the speech quality was considered, the possibility of improving the recognition performance of ASR still remains. Recently, the RL began to be introduced to train the ASR system [14, 15]. As is well known, the deep Q-network (DQN) is an effective way to implement RL, which is an area of machine learning concerned with how software agents ought to take actions in a way to maximize the reward under given environments.

In this paper, we design the DQN-based speech enhancement in such a way that the speech recognition accuracy is improved under various noise conditions. We first develop the SpinalNet to estimate noise suppression gain vectors, which are quantized to produce the finite action-value functions. The DQN is then designed based on the finite action to value the DQN higher than the SpinalNet when the DQN-based action outperforms the SpinalNet-based one in terms of the given reward. To validate the performance comparison between the DQN and SpinalNet, the ASR performance in terms of word error rate (WER) as well as the objective speech quality measure, called composite measure [16], are exploited to design the reward when training the DQN. In the test stage, the best action-value function, namely, one of the quantized noise suppression gain vectors is chosen by the DQN on a frame-by-frame basis and applied to noisy speech, leading to higher speech ASR performance.

The rest of this paper is processed in the following order. Deep Q-network is reviewed in Section 2. The proposed algorithm is introduced in Section 3. Experimental setup and results are provided in Section 4. In the end, conclusions are provided in Section 5.

2. Review of deep Q-network

The approach we present in this paper is focused on RL, a method of learning how to select an optimal action after constructing a policy which is taken to the next state in order to receive the highest reward in a specific state. Policies define how learning agents take at specific times. Specifically, a policy is an action to be taken when in that state in a detected environmental state.

In certain cases, a policy may be in the form of a lookup table or a very simple function, while in some cases it may include something like a search process. Policies are the core parts of reinforcement learning, because they are effective in determining behavior. The goals of reinforcement learning problems are defined as rewards. At each time step, the reinforcement learning agent receives a single number, a reward from the environment. Maximizing the total reward is a very important goal of the agent. The agent is informed of what is good from the reward. The reward is a fundamental component of policy change. If a low reward is given for the action chosen as a policy, the policy will be changed so that a different action is chosen in the same situation in the future. Usually, the reward can be a stochastic tool of environmental conditions and actions. Optimal actions represent good things in the long term and rewards represent good things in the relatively short term. The value of a state is the total reward that an agent can accumulate from present to future. It is originally inspired by behavioral psychology as an area of machine learning where an agent defined given an environment selects an action that maximizes the reward given the choice of the action on a current state to be recognized. These problems are so comprehensive that they are also studied in areas such as game theory, control theory, multiagent systems and so forth.

The optimization control theory also studies similar problems, but it clearly differs from the reinforcement learning approaches in terms of learning and approximation in that most studies focus on the existence and characteristics of the optimal solution. The RL also differs from general supervised learning in that there is no training set consisting of input and output pairs. This is why no explicit correction is made for wrong behavior.

Instead, the focus of the RL lies on performance in the learning process which is enhanced by balancing exploration and exploitation. Balanced problem of exploration and exploitation is the most studied problem in the RL, and it has been studied in multiarmed bandit problem and finite Markov decision process [9].

There exist several types of strategies to make use of the RL with a goal of finding the best action. A simple but efficient way is to create a Q-table where we calculate the maximum expected future reward, for each action of each state. Then, since we know what the best action to take is for each state, we are able to take that action of that state with the best policy given. Instead of implementing a policy directly, we aim at improving the Q-table to always choose the best action.

The values for each element of the Q-table are specifically updated via the Q-learning algorithm. Once the Q function (or action-value function) is defined by the expected future reward of that action at the state, the Bellman equation [17] is used to iteratively update the Q function. Then, as we explore the environment, Q gives us a better and better approximation even in a harder environment. However, when imaging a big environment with a gigantic state space producing different states, it is not easy to create and update the Q-table at all. The best idea is to use a neural network, which classifies, given a state, the different Q-values for each action. This algorithm is called the DQN to take an input and pass it through its network for finding a vector of Q-values for each action possible in the given state. Then, it is needed to take the biggest Q-value of this vector, which corresponds to the best action.

A pioneered work, presented in [11], investigated whether the RL, which is implemented on the DQN algorithm, can self-optimize a DNN-based speech enhancement algorithm. To implement the DQN, finite action functions are required for which the DNN-based enhancements gains are first quantized. Then, the DQN is carefully designed to select the best action, gain vector to suppress noise by utilizing quantitative metrics reflecting a subjective individual perception score as a reward of the RL.

Thus, the DQN exhibits the superior performance compared to the DNN-based references over a diverse range of scenarios. However, it is useful only to enhance the speech quality but does not mean that the accuracy

of speech recognition systems is improved in difficult acoustical environments.

3. Proposed DQN-based noise suppression method

In order to develop the DQN-based noise suppression for improving the ASR performance, we first find the relevant reward in measuring performance of the ASR. Since the perceptual evaluation of speech quality used in [11] does not exhibit inherently high correlation with the ASR accuracy, the first strategy is to use the WER, designed by

$$WER = \frac{S + I + D}{N} \times 100, \tag{1}$$

where the word sequence hypothesized by the ASR system is aligned with a reference transcription, and the number of errors is computed as the sum of substitutions (S), insertions (I), and deletions (D) with N denoting the number of total words. Hence, the WER becomes the Levenshtein distance between the hypothesis $\mathbb{H}$ and correct answer phase $\mathbb{R}$, namely $\mathcal{D}(\mathbb{H}, \mathbb{R})$. For this, let m be the number of words in $\mathbb{R}$ and n be the number of word in $\mathbb{H}$

$$\mathcal{D}_{i,j} = \min \begin{cases} \mathcal{D}_{i-1,j+1} + 0 & (\mathbb{H}_j = \mathbb{R}_i) \\ \mathcal{D}_{i-1,j+1} + 1 & (\text{Substitution}) \\ \mathcal{D}_{i,j-1} + 1 & (\text{Insertion}) \\ \mathcal{D}_{i-1,j} + 1 & (\text{Deletion}) \end{cases} \tag{2}$$

where $1 \leq i \leq m$, $1 \leq j \leq n$, $\mathbb{H}_j$ denotes the jth word in the hypothesis $\mathbb{R}_i$ denotes the ith word in the reference. Then, the WER is finally computed as follows:

$$\texttt{WER}(\mathbb{H}, \mathbb{R}) = \mathcal{D}_{i,j}/m, \tag{3}$$

where we believe that `WER` can be successfully responsible for the short-term recognition accuracy as the desired reward, additionally, we consider the composite measure $\mathbb{C}$, proposed in [16], which is known to show the significantly high correlation with the listening quality as perceived by users which is determined by the regression technique as shown below:

$$\mathbb{C} = 1.594 + 0.805 \cdot S_{PESQ} - 0.512 \cdot S_{LLR} - 0.007 \cdot S_{WSS}, \tag{4}$$

where S_{PESQ}, S_{LLR}, and S_{WSS} mean the values obtained by perceptual evaluation of speech quality (PESQ) [18], log likelihood ratio, and weighted spectral slope, respectively. We believe that $\mathbb{C}$ can be responsible for the long-term recognition accuracy as the desired reward. Since the composite measure $\mathbb{C}$ is used to determine the quality of the whole speech, it is wise to say that $\mathbb{C}$ accounts for the long-term reward. Then, the reward functions `WER` and $\mathbb{C}$ are collpased into the single reward, in which we use the difference of `WER` and $\mathbb{C}$ compared to opponents rather than using original values themselves, since `WER` or $\mathbb{C}$ are affected not only by the performance but also the external factors [16]. As for the opponent, the SpinalNet [8] is employed due to its superior performance. The configuration of the SpinalNet is composed of an input row, an intermediate row of multiple hidden layers, and the output row. To minimize the number of multiplications, Kabir et al., kept the number of inputs per layer and the number of neurons per hidden layers as small as possible but this may cause the network to underfit. To overcome this issue, each layer in the intermediate row receives the input from the previous layer. Since the input is recurring, if any significant input feature does not affect the output in one of the hidden layers, they can affect in another hidden layer. The input is split into two rows as depicted

in Figure 1 and both rows are allocated to different hidden layers repetitively. The intermediate row contains a nonlinear activation function and the output row contains the linear activation function. Finally, the last layer, output layer, combines the weighted outputs of the hidden neurons on the intermediate row. Thus, we obtain the below with the scaling factor α_1 and α_2:

$$\mathbb{U} = \tanh(\alpha_1(\mathtt{WER}^{DQN} - \mathtt{WER}^{SpinalNet}) + \alpha_2(\mathbb{C}^{DQN} - \mathbb{C}^{SpinalNet})), \tag{5}$$

where $\mathbb{C}^{DQN}$ and $\mathbb{C}^{SpinalNet}$ denote the composite measure values directly obtained by the DQN and SpinalNet, respectively. Also, $\mathtt{WER}^{DQN}$ and $\mathtt{WER}^{SpinalNet}$ are determined by not only the DQN and SpinalNet but also the ASR algorithm.

Next, $\mathbb{U}$ is further revised to take full consideration on the logarithmic difference between enhanced speech and clean speech as given by

$$\tilde{E}_t = \sum_{\tau=-M}^{M} \sum_{w=1}^{\Omega} |\log|\hat{X}_{w,t+\tau}| - \log|X_{w,t+\tau}||^2 \tag{6}$$

where $\hat{X}_{w,t+\tau}$ denotes an enhanced speech signal and $X_{w,t+\tau}$ denotes a clean speech signal. $\omega = \{1, 2, \cdots, \Omega\}$, $t = \{1, 2, \cdots, T\}$ and $\tau = \{0, 1, \cdots, M\}$ denote the indices of the frequency, time and context window, respectively. Subsequently, E_t is normalized between 0 and 1 such that

$$E_t = \frac{\tilde{E}_t}{\max_{t \in T}(\tilde{E}_t)}. \tag{7}$$

Then, the new reward r_t is calculated during speech is absent such that

$$r_t = \begin{cases} (1 - E_t)\mathbb{U} & (0 < \mathbb{U} < 1) \\ E_t\mathbb{U} & (-1 \leq \mathbb{U} \leq 0) \end{cases} \tag{8}$$

where the voice activity detection (VAD) algorithm, proposed in [19], was employed to identify speech absent or not. As for the speech periods, r_t is determined to exploit the ASR results obtained from the two competing algorithms, namely DQN and SpinalNet, in each time t such that

$$r_t = \begin{cases} w_t(1 - E_t)\mathbb{U} & (0 < \mathbb{U} < 1) \\ w_t E_t\mathbb{U} & (-1 \leq \mathbb{U} \leq 0) \end{cases} \tag{9}$$

In time, w_t is obtained differently for four distinct cases:

$$w_t = \begin{cases} w_1 & (\ \mathbb{H}_t^{DQN} = \mathbb{R}_t, \mathbb{H}_t^{SpinalNet} \neq \mathbb{R}_t\ ) \\ w_2 & (\ \mathbb{H}_t^{DQN} = \mathbb{R}_t, \mathbb{H}_t^{SpinalNet} = \mathbb{R}_t\ ) \\ w_3 & (\ \mathbb{H}_t^{DQN} \neq \mathbb{R}_t, \mathbb{H}_t^{SpinalNet} \neq \mathbb{R}_t\ ) \\ w_4 & (\ \mathbb{H}_t^{DQN} \neq \mathbb{R}_t, \mathbb{H}_t^{SpinalNet} = \mathbb{R}_t\ ) \end{cases} \tag{10}$$

where $\mathbb{H}_t^{DQN}$ and $\mathbb{H}_t^{SpinalNet}$ denote the hypotheses derived from the DQN algorithm and the SpinalNet, respectively, and $\mathbb{R}_t$ is the correct reference at time t.

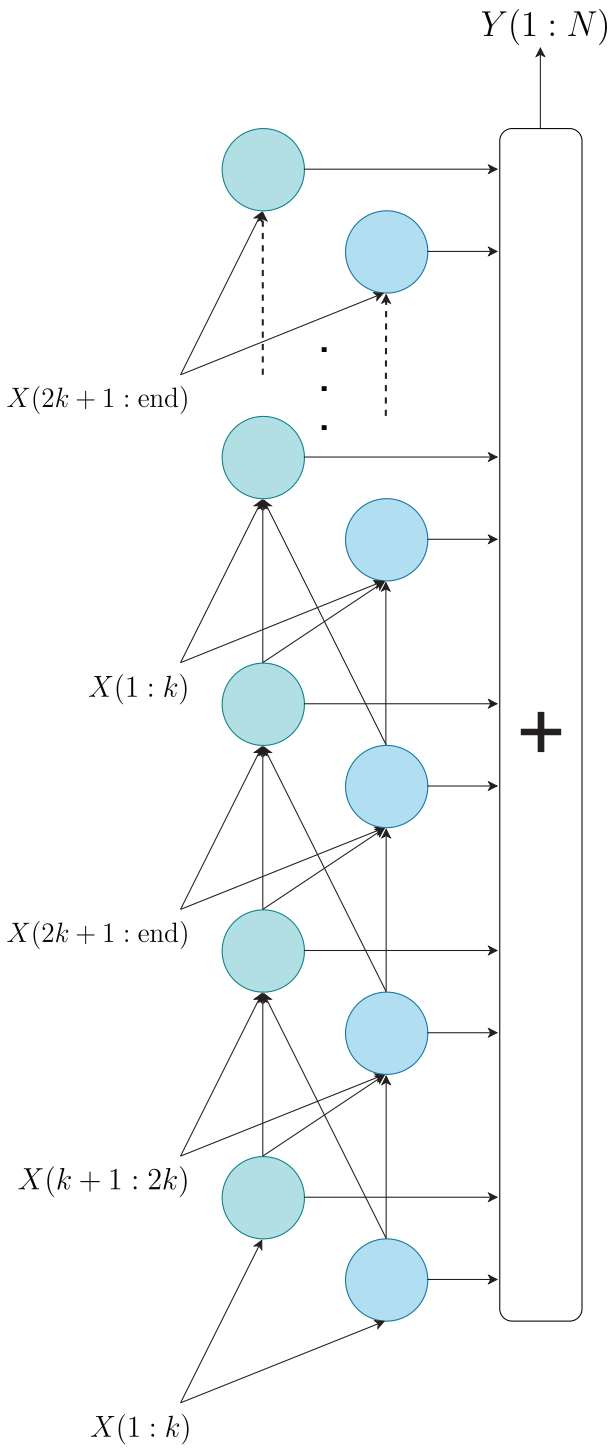

Figure 1. The structure of SpinalNet [8] as X and Y denote the input and output of network, respectively.

As shown in Figure 2, the process of calculating the reward of the proposed algorithm described above is summarized as follows. The log power spectrum (LPS) is extracted from the training dataset and it is applied

as the input to the DQN. Then, the noise suppression gain, which is the output of the DQN, is applied to the input LPS. In addition, the noise suppression gain, which is obtained through the SpinalNet, is separately applied to the input LPS. Each output from the DQN and SpinalNet is applied as input to ASR and WERs are acquired. In addition, composite measures are calculated based on the output of each algorithm. After that, the rewards are determined through Eqs. (6)–(10), eventually, the DQN is trained through the reward obtained above.

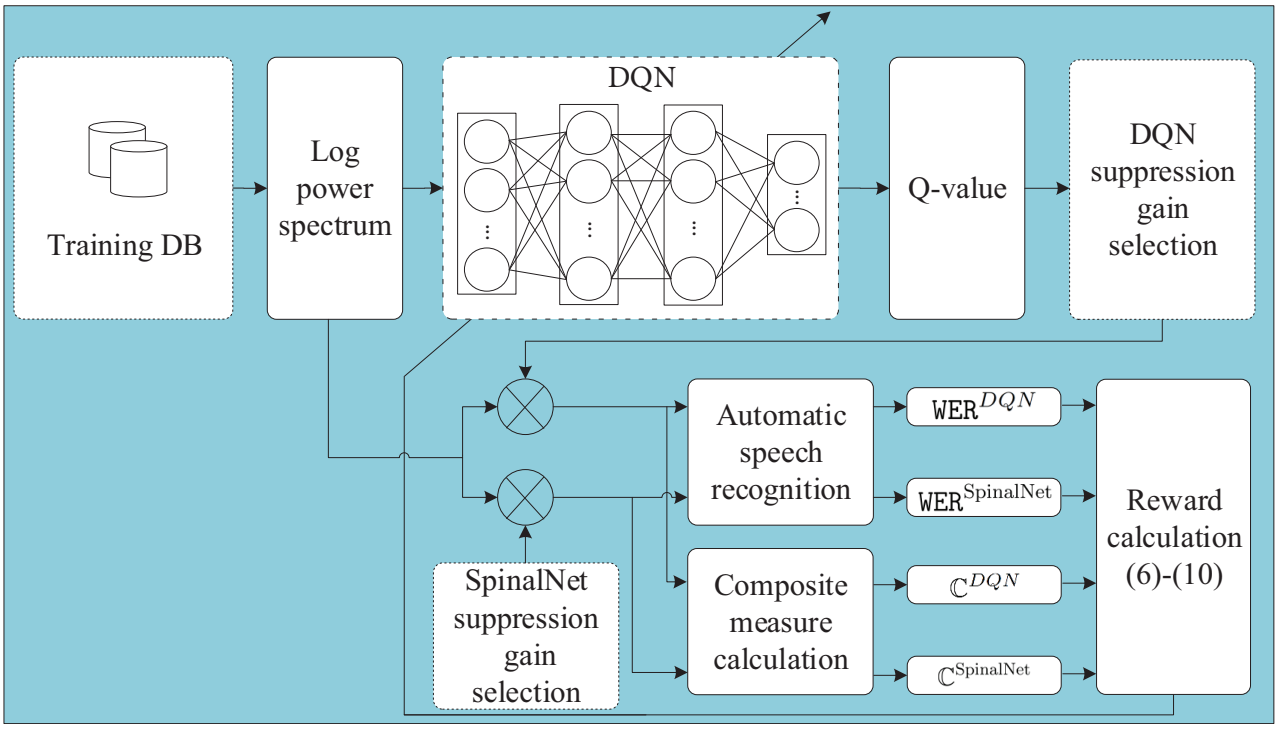

Figure 2. Overall block diagram of the proposed DQN-based noise suppression technique.

Once the reward is carefully designed, the next step is to develop the policy Q function, indicating for how good it is for the ASR system to pick an action while being in the current observation y_t. For this, the action a_t must belong to the finite set for which a continuous large set of gain functions derived from the SpinalNet noise suppression as in [8] are converted into A, discrete set with a size of $\mathbb{A}(=32)$ through the quantization step [11]. Then, the optimal selection policy $\tilde{Q}$ at the tth frame is formed to consider the DQN output in future τ frames as follows:

$$\tilde{Q}(y_t, a_t) = \begin{cases} r_t + \sum_{\tau=0}^{M} \lambda(\tau)[\max_{a \in A} Q(y_{t+\tau}, a)] & \text{if } 0 < \mathbb{U} < 1 \\ Q(y_t, a_t) & \text{otherwise} \end{cases} \tag{11}$$

where $\lambda(\tau)$ denotes the τth discount factor. Next, $\mathbb{W}_Q$, a set of parameter of the DQN is fully trained to minimize the cost, which is the error between the optimal policy $\tilde{Q}$ and the DQN output Q as follows:

$$\mathbb{W}_Q \leftarrow \arg\min_{\mathbb{W}} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{\mathbb{A}} |\tilde{Q}(y_t, i) - Q(y_t, i)|^2, \tag{12}$$

where T is the total number of the frames.

After training, in the test stage, the T-F IRM mask is determined as the best action chosen by the optimal selection policy for given input y_t as follows:

$$\hat{a}_t = \arg\max_{i \in \mathbb{A}} Q(y_t, i), \tag{13}$$

where the suppression gain $G_{w,\hat{a}_t}$ is applied to input y_t in the end such that

$$\hat{X}_{w,t} = G_{w,\hat{a}_t} \cdot Y_{w,t} \tag{14}$$

yielding the enhanced speech, which serves as the input to the ASR system.

4. Experiments and results

This section presents the performance evaluation of the proposed DQN-based noise suppression method. In order to objectively verify the superiority of our proposed method, we compared it with the conventional algorithms.

The first and second comparative models were proposed in [6, 8], respectively. The third comparative model was the conventional RL and the structure was designed as the 257 dimensions LPS corrupted by noise [11]. The frame considered were 5 previous frames, the present frame, and 5 future frames, total dimension was 2827 (= 257 dimensions $\times$ 11 frames). The hidden layer consisted of three. Each hidden layer consisted of 1024 nodes and the output was 32 dimensions, which makes the DQN results in lower complexity than the SpinalNet whose output dimension was 257. In other words, we set 32 noise suppression gain clusters established by the k-means algorithm. We used the ReLU [20] as the activation function, mean squared error (MSE) as the cost function, Adam as the optimizer [21] and 0.0005 as the learning rate. The fourth comparative model was proposed in [12].

The input dimension of the proposed DQN-based noise suppression algorithm was 2827 (= 257 dimensions $\times$ 11 frames), which is considered as a total of 11 frames, 5 previous frames, the present frame, and 5 future frames considered as a basic unit of 257 dimensional LPS of speech signal corrupted by noise. The hidden layer consisted of three. Each hidden layer consisted of 1024 nodes and the output was 32 dimensions. In other words, we set noise suppression gain vectors 32 clusters given by k-means algorithm. We used the same activation function, optimizer and learning rate as used in the third comparative model [11]. In (5), $\alpha_1 = -1, \alpha_2 = 20$ were used. Also, M was chosen 2 for in (6) and (11). In (10), each parameter was used as follows: $\omega_1 = 1.0, \omega_2 = 0.3, \omega_3 = 0.3$ and $\omega_4 = 0$. In (11), $\tau(0) = 0.7, \tau(1) = 0.2$ and $\tau(2) = 0.1$ were used. These parameters were selected with the CHiME-3 development-set [22].

The part of experiments used the CHiME-3 dataset. All the dataset used in this study have been experimented by sliding them with 512 samples and 50% overlapping at a sampling frequency of 16 kHz.

We eventually conducted experiments on the following platform. The central processing unit (CPU) was Intel i7-9700K, the random access memory (RAM) was 32 GB, and the graphics processing unit (GPU) was NVIDIA GeForce RTX 2080 Ti. The recognizer was based on Kaldi-toolkit [23] and the recognizer was trained on the CHiME-3 training-set. Specifically, the features were up to the delta-delta, which is the difference between the 13th order mel-frequency cepstrum coefficients (MFCCs) [23], namely delta and its difference, which is delta-delta, and the considered frame was 11 frames. A total of 429 (= 13 orders $\times$ 3 types $\times$ 11 frames) dimensions were considered. As for the acoustic model, the number of input nodes was 429, the hidden layer

consisted of 5, each hidden layer consisted of 1024 nodes, the activation function was sigmoid, the number of output nodes was 6063, and neutral-gradient-based optimization was applied [23]. The language model used tri-gram for 8666 words and training dataset was used about 18 hours among CHiME-3 dataset.

4.1. Experiments on CHiME-3 dataset

The dataset used in the experiment was CHiME-3, which was collected directly through a multimic mobile apparatus for ASR in environments such as bus, cafe, pedestrian region, and streets. Each dataset had two types: real speech dataset (REAL) and simulated speech dataset (SIMU). The real speech dataset was recorded in 6 channels and its sampling rate was 16 kHz. Twelve American English speakers (6 males and 6 females) aged 20–50 read sentences from the WSJ0 dataset and the sentences were recorded by a multimic mobile apparatus. As a result, the estimated signal-to-noise ratio (SNR) of the collected dataset was about 5 dB [22].

Training, development and evaluation sets were included in the CHiME-3 dataset. The first set consisted of 87 speakers, 18 h, 1600 utterances and the ratio of REAL:SIMU is 1:5, the second set consisted of 4 speakers, 2.9 h, 3280 utterances and the ratio of REAL:SIMU is 1:1. Finally, the third set consisted of 4 speakers, 2.2 h, 2640 utterances and the ratio of REAL:SIMU is 1:5. Only the fifth microphone was used in experiments.

The performance for each given noise environment was evaluated by the recognition rate in which Table 1, each value was the `WER` as summarized in Table 1. Tables 2 and 3 show the average PESQ [18] and short-time objective intelligibility (STOI) [24], respectively. A high PESQ (between –0.5 and 4.5) and STOI (between 0 and 1) indicate improved quality and intelligibility, respectively. Experimental results show that the proposed approach leads to improvements of speech recognition, PESQ, and STOI in various noisy conditions while reducing the computational burden compared to the SpinalNet.

Table 1. Performance comparison of WER for the conventional algorithms and proposed technique on CHiME-3 dataset.

Conditions		WER					
Environments	Types	Unprocessed	Arslan [6]	Kabir [8]	Koizumi [11]	Shen [12]	Proposed method
Bus	SIMU	56.09	25.12	24.08	22.61	21.74	18.64
	REAL	65.37	36.31	34.25	33.30	32.34	29.04
Cafe	SIMU	64.01	36.93	32.92	31.39	30.29	27.59
	REAL	71.85	42.74	38.17	36.88	35.40	32.16
Pedestrian	SIMU	64.65	30.76	29.29	27.54	25.77	22.27
	REAL	56.38	26.32	25.44	24.83	23.06	19.96
Street	SIMU	62.47	33.87	31.86	30.13	28.42	25.22
	REAL	55.55	25.62	23.33	21.76	19.98	16.28
Overall		62.04	32.21	29.91	28.55	27.13	23.89

From these results, it was found that RL optimized the DQN-based noise suppression algorithm, and the ASR performance of the proposed method was improved when exploiting the WER and composite measure as reward for DQN. In summary, it makes sense to say that the proposed algorithm efficiently used the training data resources which were relatively insufficient in quantity than the SpinalNet, since the DQN-based method outperformed the SpinalNet in terms of WER, PESQ, and STOI. Also, in the test phase after all training is complete, the proposed technique is to choose one of 32 mask templates per frame. In contrast, the SpinalNet does a regression with 257 dimensions of output per frame. Therefore, the complexity of the proposed method is

Table 2. Performance comparison of PESQ for the conventional algorithms and proposed technique on CHiME-3 dataset.

Conditions		WER					
Environments	Types	Unprocessed	Arslan [6]	Kabir [8]	Koizumi [11]	Shen [12]	Proposed method
Bus	SIMU	0.82	1.39	1.41	1.44	1.46	1.52
	REAL	0.65	1.18	1.22	1.24	1.26	1.33
Cafe	SIMU	0.67	1.17	1.25	1.28	1.30	1.35
	REAL	0.53	1.06	1.15	1.17	1.21	1.27
Pedestrian	SIMU	0.66	1.29	1.31	1.35	1.38	1.45
	REAL	0.81	1.37	1.38	1.40	1.44	1.50
Street	SIMU	0.70	1.23	1.27	1.30	1.34	1.40
	REAL	0.83	1.38	1.42	1.46	1.49	1.56
Overall		0.71	1.26	1.30	1.33	1.36	1.42

Table 3. Performance comparison of STOI for the conventional algorithms and proposed technique on CHiME-3 dataset.

Conditions		WER					
Environments	Types	Unprocessed	Arslan [6]	Kabir [8]	Koizumi [11]	Shen [12]	Proposed method
Bus	SIMU	0.306	0.519	0.527	0.538	0.546	0.568
	REAL	0.242	0.442	0.456	0.464	0.472	0.495
Cafe	SIMU	0.251	0.437	0.466	0.477	0.486	0.506
	REAL	0.196	0.397	0.429	0.439	0.450	0.474
Pedestrian	SIMU	0.247	0.480	0.491	0.504	0.517	0.543
	REAL	0.304	0.511	0.517	0.523	0.536	0.559
Street	SIMU	0.262	0.459	0.473	0.486	0.499	0.522
	REAL	0.310	0.516	0.532	0.544	0.558	0.584
Overall		0.265	0.470	0.486	0.497	0.508	0.531

significantly smaller than that of SpinalNet. The utilization of WER and composite measure, which are closely related to speech recognition performance, to the reward design can be seen as a factor contributing to the performance difference between the proposed algorithm and conventional reinforcement learning methods.

5. Conclusion

In this paper, we proposed the DQN-based noise suppression algorithm which was designed by the reward based on the WER and composite measure exhibiting higher correlation with ASR performance by referring to existing speech enhancement based on RL whose reward is PESQ. The proposed method was evaluated in terms of WER, PESQ, and STOI conducted on CHiME-3 dataset and outperformed conventional methods under various noise conditions.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System). Joon-Hyuk CHANG gave the idea, Tae-Jun

PARK did the experiments, Tae-Jun PARK and Joon-Hyuk CHANG interpreted the results, Tae-Jun PARK wrote the paper.

References

[1] Xu Y, Du J, Dai LR, Lee CH. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech and Language Processing 2015; 23 (1): 7-9. doi: 10.1109/TASLP.2014.2364452

[2] Narayanan A, Wang DL. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: IEEE 2013 International Conference on Acoustics, Speech, and Signal Processing; Vancouver, Canada; 2013. pp. 7092-7096.

[3] Deng L, Yu D, Hinton G. Deep learning for speech recognition and related applications. In: Annual Conference on Neural Information Processing Systems 2009; Vancouver, Canada; 2009.

[4] Jin Z, Wang D. A supervised learning approach to monaural segregation of reverberant speech. IEEE Transactions on Audio, Speech, and Language Processing 2009; 17 (4): 625-638. doi: 10.1109/TASL.2008.2010633

[5] Williamson DS, Wang Y, Wang DL. Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2016; 24 (3): 483-492. doi: 10.1109/TASLP.2015.2512042

[6] Arslan Ö, Engín EZ. Speech enhancement using adaptive thresholding based on gamma distribution of Teager energy operated intrinsic mode functions. Turkish Journal of Electrical Engineering & Computer Sciences 2019; 27 (2): 1355-1370. doi: 10.3906/elk-1804-18

[7] Zhao H, Zarar S, Tashev I, Lee CH. Convolutional-Recurrent Neural Networks for Speech Enhancement. In: IEEE 2018 International Conference on Acoustics, Speech, and Signal Processing; Calgary, Canada; 2018. pp. 2401-2405.

[8] Kabir H, Abdar M, Jalali SMJ, Khosravi A, Atiya AF et al. SpinalNet: Deep neural network with gradual input. arXiv 2020. arXiv:2007.03347v2

[9] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J. Human-level control through deep reinforcement learning. Nature 2015; 518: 529-533. doi: 10.1038/nature14236

[10] Sugiyama M. Statistical Reinforcement Learning: Modern Machine Learning Approaches. Boca Raton, FL, USA: Chapman and Hall/CRC, 2015.

[11] Koizumi Y, Niwa K, Hioka Y, Kobayashi K, Haneda Y. DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements. In: IEEE 2017 International Conference on Acoustics, Speech, and Signal Processing; New Orleans, USA; 2017. pp. 81-85.

[12] Shen Y, Huang C, Wang S, Tsao Y, Wang H et al. Reinforcement learning based speech enhancement for robust speech recognition. In: IEEE 2019 International Conference on Acoustics, Speech, and Signal Processing; Brighton, UK; 2019. pp. 6750-6754.

[13] Fakoor R, He X, Tashev I, Zarar S. Reinforcement learning to adapt speech enhancement to instantaneous input signal quality. In: Annual Conference on Neural Information Processing Systems 2017; Long Beach, CA, USA; 2017.

[14] Kala T, Shinozaki T. Reinforcement learning of speech recognition system based on policy gradient and hypothesis selection. In: IEEE 2018 International Conference on Acoustics, Speech, and Signal Processing; Calgary, AB, Canada; 2018. pp. 5759-5763.

[15] Hori T, Astudillo R, Hayashi T, Zhang Y, Watanabe S et al. Cycle-consistency training for end-to-end speech recognition. In: IEEE 2019 International Conference on Acoustics, Speech, and Signal Processing; Brighton, UK; 2019. pp. 4723-4725.

[16] Hu Y, Loizou P. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing 2008; 16 (1): 229-238. doi: 10.1109/TASL.2007.911054

[17] Bellman RE. Dynamic Programming. Princeton, NJ, USA: Princeton University Press, 1957.

[18] ITU-T, Rec. P.862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union-Telecommunication Standardization Sector 2001.

[19] Zhang XL, Wang D. Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2016; 24 (2): 252-264. doi: 10.1109/TASLP.2015.2505415

[20] Agarap AF. Deep learning using rectified linear units (relu). arXiv 2018. arXiv:1803.08375.

[21] Kingma D, Ba J. Adam: a method for stochastic optimization. In: International Conference on Learning Representations 2015; San Diego, CA, USA; 2015.

[22] Barker J, Marxer R, Vincent E, Watanabe S. The third CHiME speech separation and recognition challenge: dataset, task and baselines. In: IEEE 2015 Workshop on Automatic Speech Recognition and Understanding; Scottsdale, AZ, USA; 2015.

[23] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O et al. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding; Waikoloa, HI, USA; 2011.

[24] Taal CH, Hendriks RC, Heusdens R, Jensen J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proceedings of the IEEE 2010 International Conference on Acoustics, Speech and Signal Processing; Dallas, TX, USA; 2010. pp. 4214–4217.