DNN-Based Monaural Speech Enhancement with Temporal and Spectral Variations Equalization

Tae Gyoon Kang[a,*], Jong Won Shin[b,**], Nam Soo Kim[a]

[a]*Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, Seoul, Korea*
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-8439

[b]*School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Korea*
123 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Korea
Tel: +82-62-715-2235

Abstract

Recently, deep neural networks (DNNs) were successfully introduced to the speech enhancement area. Conventional DNN-based algorithms generally produce over-smoothed output features which deteriorate the quality of the enhanced speech. In addition, their performance measures calculated in the linear frequency scale do not match the human auditory perception where the sensitivity follows the Mel-frequency scale. In this paper, we propose a novel objective function for DNN-based speech enhancement algorithm. In the proposed technique, a new objective function which consists of the Mel-scale weighted mean square error, and temporal and spectral variations similarities between the enhanced and clean speech is employed in the DNN training stage. The proposed objective function helps to compute the gradients based on a perceptually motivated non-linear frequency scale and alleviates the over-smoothness of the estimated speech. In the experiments, the performance of the proposed algorithm was compared to the conventional DNN-based speech enhancement algorithm in matched and mismatched noise conditions. From the experimental results, we can see that the proposed algorithm performs better than the

[*] This author is currently with the Samsung Electronics.

[**] Corresponding author

Email address: jwshin@gist.ac.kr (Jong Won Shin)

conventional algorithm in terms of both the objective and subjective measures.

Keywords: Monaural speech enhancement, deep neural network, objective function

1. Introduction

For a few decades, monaural speech enhancement from additive noise signal has been widely studied to improve various communication and signal processing systems [1, 2]. Though considerable performance improvements have been achieved by various approaches, speech enhancement in realistic noise environments still remains a challenging problem.

Early studies on monaural speech enhancement are mostly based on the minimum mean-square error (MMSE) criterion [3, 4] which has improved the perceptual speech quality with an affordable amount of musical noise. The quality of the enhanced speech was further improved by adopting various techniques estimating minimum statistics of the acoustic environments minima controlled recursive averaging noise estimation [5, 6, 7, 8]. However, algorithms based on this approach have difficulties in tracking non-stationary or speech-like noises which cause speech quality degradation in real-world applications.

In order to enhance the noisy speech in various noise environments, deep neural networks (DNNs) which can learn complicated inter-dependencies between the input variables [9, 10, 11, 12, 13] were successfully introduced to the speech enhancement area [14, 15, 16]. In these approaches, the DNN provides a mapping between consecutive noisy speech frames and the corresponding clean speech frame with its deep hidden structure. Furthermore, in [17], global variance (GV) equalization post-filter, dropout training, and noise-aware training techniques were incorporated to DNN-based speech enhancement to improve the speech quality in mismatched noise conditions.

Many studies have applied the DNN-based approaches to speech enhancement and target speaker separation with various new ideas. Huang et al. proposed a technique to jointly optimize all the sources with a discriminative objective function for DNN and recurrent neural network (RNN) [18]. Han et al. applied a DNN-based method for joint dereverberation and denoising followed by iterative signal reconstruction [19]. The training targets of the DNNs were studied in [20], and the phase-sensitive filter and complex ratio masking were also proposed [21, 22]. Zhang et al. investigated the performance of the mapping- and masking-based training targets both theoretically and experimentally in [23] where they also proposed the multi-context stacking networks for deep ensemble learning. The multi-objective learning scheme was adopted to utilize secondary targets in [24]. Finally, the divide and conquer strategy was applied by hierarchical DNN and SNR-based progressive learning algorithms [25, 26].

Conventional DNN-based speech enhancement algorithms generally apply the objective functions which are related to the mean square error between the enhanced and clean speech features [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. Since these measures compute the errors from various frequency bins with linear frequency scale, they do not align with the human auditory perception where the sensitivity follows the Mel-frequency scale. The perceptual quality of the enhanced speech would be improved if the cost function of DNN reflects relative importance of frequency components considering this nonlinear frequency sensitivity.

In addition, it is well-known that the estimated speech trajectories obtained from the DNN-based algorithms are usually over-smoothed compared to those of the clean speech, since the conventional mean square error measures are derived from each time-frequency bin separately rather than from whole spectral trajectory [17]. The speech generated from these enhancement algorithms may result in muffled sound quality and decreased intelligibility [17, 27, 28]. Several studies have applied the element-wise weight function and the penalty term to the conventional mean square error [29, 30]. However, their works were not closely related to the human auditory perception.

In this paper, we propose a novel DNN-based speech enhancement algorithm which computes the gradients based on a perceptually motivated non-linear frequency scale and alleviates the over-smoothness problem by equalizing temporal and spectral variations of the enhanced speech to match those of the clean speech. The main contributions of the proposed algorithm are summarized as follows:

First, we apply the Mel-scale weight to fit the objective function to the critical frequency bands of hearing. Similar to the human auditory perception, the network trained using the Mel-scaled gradients is more sensitive to the perceptually important frequency bins. The Mel-frequency scale was adopted to speech enhancement in [31] to smooth the gain function over spectral coefficients. In contrast, the Mel-scale is introduced to prioritize the gradients according to the perceptual importance in the proposed algorithm.

Second, the objective function for DNN training is modified to incorporate the temporal and spectral variation similarities between the enhanced and clean speech. By equalizing the temporal and spectral variations, the enhanced speech could have the spectral peaks and valleys distributed similarly to those of the clean speech. The proposed objective functions are motivated by the relation between the human intelligibility and short-time analysis on one-third octave band trajectory [32]. We adopt variation similarity over short-time trajectories and spectral coefficients into the DNN-based speech enhancement framework and analyze their effect on the naturalness and intelligibility of the enhanced speech. While the long short-term memory (LSTM) [33] and gated recurrent unit (GRU) [34] can learn the temporal relations of the consecutive input frames, they cannot directly compensate the lack of output feature structure as the proposed approach, and thus these approaches can be jointly applied [35].

The rest of this paper is organized as follows: an overview of the conventional DNN-based speech enhancement technique is given in Section 2 and the Mel-scale weighted mean square error and variation similarities are described in Section 3. Then, the performance evaluation of DNNs with various training algorithms are provided in Section 4. Finally, conclusions are drawn in Section 5.

2. Conventional DNN-based speech enhancement

The task of DNN-based speech enhancement can be divided into the training and test stages. In the training stage, the noisy speech features and the corresponding clean speech features are respectively fed to the input and output nodes of the DNN, and the network is optimized to minimize the mean square error between the enhanced and clean speech features. After the training stage, the clean speech features are estimated from the noisy speech features through the DNN and a GV equalization post-filter is applied to compensate the over-smoothed output trajectory. In this section, we present the feature structures and training scheme of the conventional DNN-based speech enhancement algorithm.

2.1. Training Stage

In the training stage, the input and output features of the DNN are respectively extracted from the noisy speech utterances and corresponding clean speech utterances. The input and output features of the DNN are usually normalized to have zero mean and unit variance before being fed to the network.

For the input and output features, we extract log-power spectra of the noisy and clean speech as in [17, 19, 36]. Recent studies have compared the performance of the mapping-based method which directly estimates the clean speech to the masking-based method which produces the binary or ratio mask targets [20, 23, 36, 37]. It is controversial which method results in better performance [23, 36]. Although this paper focuses on the mapping-based method, the proposed algorithm can also be applied to the masking-based method with slight modification to generate the masked clean speech features.

Let us denote F-dimensional normalized log-power spectra of the noisy speech and clean speech at the t-th frame as $\mathbf{z}_t$ and $\mathbf{y}_t$, respectively. Then, the input feature vector $\mathbf{v}_t^0$ is generally constructed as follows:

$$\mathbf{v}_t^0 \quad = \quad [\mathbf{z}_{t-K}^{\dagger},\ \mathbf{z}_{t-K+1}^{\dagger},\ \cdots\ ,\ \mathbf{z}_{t+K}^{\dagger}]^{\dagger} \qquad (1)$$

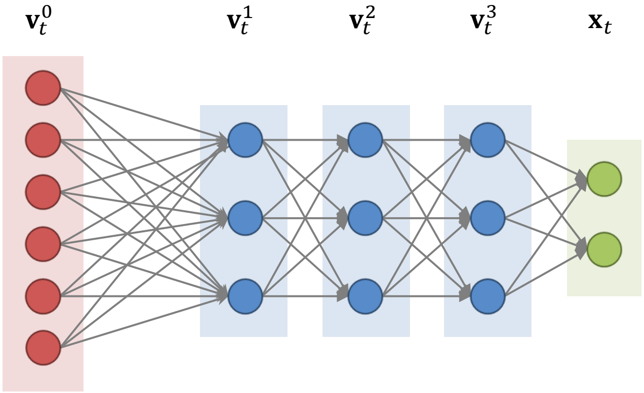

Figure 1: Scheme of the DNN with three hidden layers.

where K denotes an input context expansion parameter and $\mathbf{z}_t^\dagger$ denotes the transpose of a vector $\mathbf{z}_t$.

Fig. 1 shows the structure of a typical DNN with three hidden layers. The DNN consists of an input layer, a few hidden layers and an output layer which are fully connected to their adjacent layers. For the sake of notational simplicity, the number of hidden layers is assumed to be L and the input and output layers of the DNN are denoted as the 0-th and $(L+1)$-th layers of the DNN, respectively.

The number of nodes in the l-th layer is denoted by n_l. The n_l-dimensional activation vector $\mathbf{v}_t^l$ is generated as

$$\mathbf{v}_t^l = g(\mathbf{a}_t^l) = g(W^l \mathbf{v}_t^{l-1} + \mathbf{b}^l) \tag{2}$$

where $\mathbf{a}_t^l$, W^l, and $\mathbf{b}^l$ denote the n_l-dimensional pre-activation vector, $n_l \times n_{l-1}$-dimensional weight matrix and n_l-dimensional bias vector, respectively, and $g(\cdot)$ represents an element-wise activation function. In this paper, all hidden layers of the DNN are assumed to use the rectified linear function which can prevent the vanishing gradient problem [38], i.e.,

$$g(\mathbf{a}_t^l(i)) = max(\mathbf{a}_t^l(i), 0) \tag{3}$$

where $\mathbf{a}_t^l(i)$ denotes the i-th element of a vector $\mathbf{a}_t^l$.

After all the hidden layer activations are computed, the F-dimensional output vector $\mathbf{x}_t$ is produced by

$$\mathbf{x}_t = W^{L+1}\mathbf{v}_t^L + \mathbf{b}^{L+1}. \tag{4}$$

In this paper, the parameters of the DNN are initialized randomly [39] and optimized using the stochastic gradient descent algorithm. In the training stage, the mean square error between the network output $\mathbf{x}_t$ and the target feature $\mathbf{y}_t$ is minimized, which is given by

$$C_{mse} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(\mathbf{x}_t(f) - \mathbf{y}_t(f))^2 \tag{5}$$

where T denotes the total number of frames of the given training data.

2.2. Test Stage

In the test stage, the clean speech estimate $\mathbf{x}_t$ is obtained from $\mathbf{v}_t^0$ through the standard feedforward processing. In the speech enhancement algorithm without GV equalization, $\mathbf{x}_t$ is de-normalized to $\bar{\mathbf{x}}_t$ as follows:

$$\bar{\mathbf{x}}_t = \mathbf{x}_t \otimes \mathbf{s} + \mathbf{m} \tag{6}$$

where $\mathbf{m}$ and $\mathbf{s}$ are respectively the mean and standard deviation vectors used to normalize the output feature of the DNN, and $\otimes$ denotes element-wise multiplication between two vectors. In this paper, only the magnitude spectrum of the speech is estimated while the phase parts of the noisy speech are kept intact.

One of the significant drawbacks of the conventional DNN-based speech enhancement algorithm is that it usually results in over-smoothed spectral trajectories of the enhanced speech. In order to alleviate this phenomenon, the GV equalization post-filter which modifies the variance of $\mathbf{x}_t$ to match that of $\mathbf{y}_t$ is usually employed. In this paper, the frequency-independent GV equalization method which has been known to perform better than the frequency-dependent approach [17] is applied as a conventional post-filtering technique.

In the frequency-independent GV equalization, the global variances of $\mathbf{x}_t$ and $\mathbf{y}_t$ are computed as follows:

$$GV(\mathbf{x}) = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(\mathbf{x}_t(f) - \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\mathbf{x}_t(f))^2, \tag{7}$$

$$GV(\mathbf{y}) = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(\mathbf{y}_t(f) - \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\mathbf{y}_t(f))^2. \tag{8}$$

Based on (7) and (8), the frequency-independent GV factor α is given by

$$\alpha = \sqrt{\frac{GV(\mathbf{y})}{GV(\mathbf{x})}}, \tag{9}$$

and it is multiplied to $\mathbf{x}_t$ before de-normalization as follows:

$$\bar{\mathbf{x}}_t = \alpha\ \mathbf{x}_t \otimes \mathbf{s} + \mathbf{m}. \tag{10}$$

In the GV equalization post-filter, multiplying the GV factor to the output feature can be viewed as imposing an exponential factor in the linear spectral magnitude domain. By this post-filter, the variance of the spectral trajectory is enlarged or diminished depending on the value of α. In most cases, α is bigger than 1 and the lack of dynamics in $\mathbf{x}_t$ is alleviated to some extent.

3. Perceptually-Motivated Criteria

In this section, we propose a novel speech enhancement algorithm that is based on DNN. We introduce the proposed objective function which consists of the Mel-scale weighted mean square error, and the temporal and spectral variation similarities between the enhanced and clean speech over adjacent frames or frequency bins.

3.1. Perceptually Motivated Objective Function

Our framework to incorporate the perceptually motivated criteria is to replace the conventional mean square error C_{mse} given in (5) by a modified objective function C defined as

$$C = \lambda_m C_{wmse} + \lambda_t(1 - \rho_{temp}) + \lambda_s(1 - \rho_{spec}) \tag{11}$$

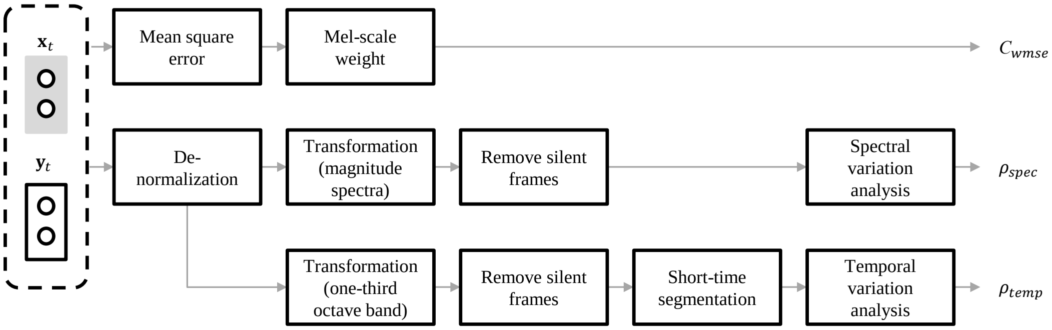

Figure 2: Scheme of the proposed objective function which incorporates Mel-scale weighted mean square error, temporal and spectral variation similarities.

where λ_m, λ_t and λ_s denote the weights controlling the contributions of the three separate sub-costs, C_{wmse}, $(1-\rho_{temp})$, and $(1-\rho_{spec})$. Fig. 2 shows the procedures for computing these three sub-costs given $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$.

In the training stage, parameters of the network are optimized so as to minimize C via the stochastic gradient descent algorithm. The test stage of the DNN remains the same to that of the conventional DNN approach. Note that the only difference of the proposed method from the conventional DNN-based speech enhancement algorithm is that it applies a new objective function for DNN training. Now, we will give the detail on how to derive C_{wmse}, ρ_{temp}, and ρ_{spec} which jointly specify the objective function. While the proposed objective function is derived from the normalized log-power spectra features, they can also be obtained from other DNN output features such as ratio masks in similar ways.

3.2. Mel-Scale Weighted Mean Square Error C_{wmse}

We modify the original mean square error C_{mse} in (5) to take the Mel-frequency scale into consideration. In this paper, the Mel-frequency scale is adopted to determine weights for the errors from various frequency bins discriminatively. The Mel-frequency is defined as follows [40]:

$$\varpi = 2595 \log_{10}(1 + \frac{\zeta}{700}) \tag{12}$$

where ϖ and ζ denote the Mel-frequency and the corresponding linear frequency, respectively. The relative importance of each spectral coefficient can be determined by the derivative of the Mel-frequency at the corresponding frequency, i.e.,

$$d(f) = min\left(\frac{d\varpi}{d\zeta}|_{\zeta=f}, \eta\right) \tag{13}$$

where η is a constant setting the minimum weight value. In (13), $d(f)$ is proportional to the frequency sensitivity of the human auditory system for the frequency f.

The Mel-scale weighted mean square error C_{wmse} is defined by multiplying the normalized weight $w(f)$ with each element of C_{mse} as follows:

$$w(f) = \frac{d(f)}{\sum_{f=1}^{F} d(f)}, \tag{14}$$

$$C_{wmse} = \frac{1}{T}\sum_{t=1}^{T}\sum_{f=1}^{F} w(f)(\mathbf{x}_t(f) - \mathbf{y}_t(f))^2. \tag{15}$$

Compared to C_{mse}, C_{wmse} emphasizes the errors in frequency bins which are more crucial for human auditory perception.

In [20], the gammatone filterbank was directly adopted to generate the target feature of the DNN. It aligns with the proposed algorithm in that human perception properties are incorporated. The proposed method employs Mel-scale weights in the objective function while the target feature remains unaltered.

3.3. Temporal Variation Similarity ρ_{temp}

It has been known that the similarity in frequency band trajectories between the enhanced and clean speech is related to the intelligibility of the enhanced speech [27, 28]. In [32], the short-time objective intelligibility (STOI) which is a speech intelligibility metric using temporal variation over the one-third octave band trajectory is presented. The one-third octave band domain was adopted since the formant fluctuation within frequency bands should be ignored for temporal variation analysis. Motivated by these studies, we attempt to equalize the temporal variation of the one-third octave band trajectory of the enhanced speech during the DNN training session.

The comparison in temporal variation between the enhanced and clean speech is performed similarly to [32]. In the DNN training stage, the output feature vectors are transformed into the one-third octave band domain before the short-time segmentation and variation analysis are performed to obtain the temporal variation similarity for each slice of frames. Then, we incorporate the temporal variation similarity values to the objective function and compute the gradients from them.

The de-normalized versions of the enhanced and clean log-power spectra $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{y}}_t$ are transformed to the H-dimensional one-third octave band vectors $\tilde{\mathbf{x}}_t$ and $\tilde{\mathbf{y}}_t$ as follows:

$$\begin{aligned} \tilde{\mathbf{x}}_t &= \sqrt{B\ exp(\bar{\mathbf{x}}_t)}, & (16) \\ \tilde{\mathbf{y}}_t &= \sqrt{B\ exp(\bar{\mathbf{y}}_t)} & (17) \end{aligned}$$

where B denotes the $H \times F$-dimensional one-third octave band matrix, and $exp(\mathbf{x})$ and $\sqrt{\mathbf{x}}$ denote the element-wise exponential and square root functions of a vector $\mathbf{x}$, respectively. The temporal variation similarity is computed only for the speech active frames. To remove the speech absence frames from the variation analysis, a simple decision rule is applied to $\tilde{\mathbf{y}}_t$ as in [32].

The variation analysis is performed for each one-third octave band and each slice of N speech active frames. Let us denote the vectors stacking the h-th one-third octave band coefficients from the t-th frame to the $t+N-1$-th frame of the enhanced and clean speech as $\check{\mathbf{x}}_{t,h}$ and $\check{\mathbf{y}}_{t,h}$. Then, the temporal variation similarity between $\check{\mathbf{x}}_{t,h}$ and $\check{\mathbf{y}}_{t,h}$ is defined as follows:

$$\rho_{temp}(t,h) = \frac{(\check{\mathbf{x}}_{t,h} - \mu_{\check{\mathbf{x}}_{t,h}}\mathbf{1}_N)^{\dagger}(\check{\mathbf{y}}_{t,h} - \mu_{\check{\mathbf{y}}_{t,h}}\mathbf{1}_N)}{||\check{\mathbf{x}}_{t,h} - \mu_{\check{\mathbf{x}}_{t,h}}\mathbf{1}_N||\ ||\check{\mathbf{y}}_{t,h} - \mu_{\check{\mathbf{y}}_{t,h}}\mathbf{1}_N||} \qquad (18)$$

where $||\cdot||$ is an l_2-norm, $\mathbf{1}_N$ denotes an N-dimensional vector with all elements being 1 and

$$\mu_{\check{\mathbf{x}}_{t,h}} = \frac{1}{N}\sum_{i=1}^{N}\check{\mathbf{x}}_{t,h}(i), \tag{19}$$

$$\mu_{\check{\mathbf{y}}_{t,h}} = \frac{1}{N}\sum_{i=1}^{N}\check{\mathbf{y}}_{t,h}(i). \tag{20}$$

The proposed objective function incorporates the variation similarity $\rho_{temp}(t,h)$ averaged over a time-frequency window as given by

$$\rho_{temp} = \frac{1}{(T-N+1)H}\sum_{t=1}^{T-N+1}\sum_{h=1}^{H}\rho_{temp}(t,h). \tag{21}$$

By training the DNN while considering this sub-cost, the short-time trajectories of the enhance speech would have temporal variation more similar to those of the clean speech.

3.4. Spectral Variation Similarity ρ_{spec}

The speech generated by the enhancement algorithms would suffer from the muffling effect when the spectral peaks and valleys are over-smoothed [17]. In order to improve the spectral dynamics of the enhanced speech, we also introduce a variation over the frequency bins, which results in a better contrast between the spectral peaks and valleys.

The spectral variation similarity ρ_{spec} is derived in a similar manner to ρ_{temp}. However, compared to ρ_{temp}, ρ_{spec} is different in two aspects. First, ρ_{spec} is derived in the spectral magnitude domain without the frequency warping, since full frequency resolution is desirable to restore spectral dynamics. Second, while ρ_{temp} considers speech trajectory and disregards the variation over different frequency bins, ρ_{spec} aims to adjust the spectral peaks and valleys in the same time frame.

The F-dimensional enhanced and clean speech magnitude spectra $\tilde{\mathbf{x}}'_t$ and $\tilde{\mathbf{y}}'_t$ are obtained as follows:

$$\tilde{\mathbf{x}}'_t \;=\; \sqrt{exp(\bar{\mathbf{x}}_t)}, \tag{22}$$

$$\tilde{\mathbf{y}}'_t \;=\; \sqrt{exp(\bar{\mathbf{y}}_t)}. \tag{23}$$

The spectral variation similarity is computed only over the speech active frames. The variation similarity $\rho_{spec}(t)$ computed at the t-th frame is given by

$$\rho_{spec}(t) = \frac{(\tilde{\mathbf{x}}'_t - \mu_{\tilde{\mathbf{x}}'_t}\mathbf{1}_F)^\dagger(\tilde{\mathbf{y}}'_t - \mu_{\tilde{\mathbf{y}}'_t}\mathbf{1}_F)}{||\tilde{\mathbf{x}}'_t - \mu_{\tilde{\mathbf{x}}'_t}\mathbf{1}_F||\ ||\tilde{\mathbf{y}}'_t - \mu_{\tilde{\mathbf{y}}'_t}\mathbf{1}_F||} \tag{24}$$

with

$$\mu_{\tilde{\mathbf{x}}'_t} = \frac{1}{F}\sum_{i=1}^{F}\tilde{\mathbf{x}}'_t(i), \tag{25}$$

$$\mu_{\tilde{\mathbf{y}}'_t} = \frac{1}{F}\sum_{i=1}^{F}\tilde{\mathbf{y}}'_t(i). \tag{26}$$

Then, ρ_{spec} is obtained by averaging $\rho_{spec}(t)$ over all frames i.e.,

$$\rho_{spec} = \frac{1}{T}\sum_{t=1}^{T}\rho_{spec}(t). \tag{27}$$

3.5. DNN Training with the Proposed Objective Function

In the training stage with the proposed objective function, the derivative of the objective function with respect to each network output $\frac{\partial C}{\partial \mathbf{x}_t(f)}$ is computed and used to derive the gradient with respect to each parameter through back-propagation. In Appendix, we provide the details on the derivation of $\frac{\partial C}{\partial \mathbf{x}_t(f)}$.

4. Experiments

In order to evaluate the performance of the proposed algorithm, we conducted experiments in matched and mismatched noise conditions. In the experiments, 4,620 utterances of clean speech data were taken from the TIMIT training database to train the DNN. The {con_mono_1, met_mono_1, off_mono_1, car_mono_1, rai_mono_1, res_mono_1, train, traffic} noises from ITU-T recommendation P.501 database [41] and the {white, factory, babble, machinegun} noises from NOISEX-92 database [42] were used for training to simulate various noise environments including more realistic conditions. Each noise waveform was re-sampled to 16 kHz, and we chose the left channel of the binaural noise recordings in ITU-T recommendation P.501 database. For each pair of the clean speech utterance and noise waveform, a noisy speech utterance was artificially generated with an SNR value randomly chosen from {-5, 0, 5, 10, 15, 20} dB. As a result, 55,440 utterances (about 47 hours) were used in total. According to [17], the size of training data barely affects the quality of enhanced speech when the training data is more than 20 hours. A 512-point Hamming window with 50% overlap was applied. K and τ were fixed to 257 and 5, respectively (feature vectors extracted from 11 consecutive frames were concatenated similarly to [17, 19]). The context expansion parameter K was decided according to [17] which showed that the context windows longer than $K = 5$ do not increase the performance.

For the test set, 30 utterances of clean speech data were taken randomly from the TIMIT test database. The {con_mono_1, res_mono_1} noises from ITU-T recommendation P.501 database and the white noise from NOISEX-92 DB were used for the experiment in matched noise conditions. For the experiment in mismatched noise conditions, the {cafeteria, kids, street} noises from ITU-T recommendation P.501 DB were chosen. For each pair of the clean speech utterance and the noise waveform, the noisy speech utterances were artificially generated with the SNR ranging from -5 to 10 dB with 5 dB step.

The DNNs were implemented using the Theano neural network toolkit [43]. The DNNs were constructed by stacking 3 hidden layers with 2048 nodes each. The numbers of the input and output nodes were $257 \times 11 = 2827$ and 257, respectively. All networks were trained through 50 epochs. The learning rate was fixed to 0.003 in the first 10 epochs and decreased by 10% after each subsequent epoch. The momentum rate was 0.5 for the first 5 epochs and increased to 0.9 afterward. The dropout rates of the input layer and all hidden layers were set to 0.1 and 0.2, respectively. Each utterance in the training data was treated as a mini-batch in the training stage. The average value of mini-batch size was 190.6, and N was fixed to 30.

In the experiments, the enhanced speech signals obtained from DNNs with various training objective functions and GV equalization post-filter were compared in both objective measures and subjective test. The performance of the

DNN-based algorithm with the proposed techniques was compared to that with the conventional mean square error and frequency-independent GV equalization post-filter [17].

The perceptual evaluation of speech quality (PESQ) score [44], STOI value [32] and the speech to distortion ratio (SDR) [45] were used as objective measures. For the subjective measures, a preference test was conducted with the enhanced speech obtained in the mismatched noise conditions.

4.1. Performance Evaluation with Various Weight Parameters

First, we evaluated how the variation similarities ρ_{temp} and ρ_{spec} affect the performance of the enhancement algorithm by varying weight parameters λ_m, λ_t and λ_s in (11). In this experiment, we measured PESQ scores and STOI values of the enhanced speech while varying λ_m, λ_t and λ_s.

Table 1: Results of PESQ scores and STOI values with various objective functions and GV equalization post-filter averaged over various SNR values and matched noise conditions.

Objective function	λ_m	λ_t	λ_s	PESQ	STOI	SDR
C_{mse} (5)	-	-	-	2.55	0.80	9.96
C (11)	1	0	0	2.64	0.81	10.83
	1	0.5	0.5	2.72	0.84	11.41
	1	1	1	2.76	0.84	11.50
	1	2	2	2.78	0.85	11.54
	1	5	5	**2.80**	0.85	**11.56**
	1	10	10	**2.80**	**0.86**	11.50
C (11)	0	5	5	2.49	0.84	7.82

Table I shows the PESQ scores and STOI values averaged over all SNR values and noise types in the matched noise conditions. The results show that both the PESQ scores and STOI values gradually increased as λ_t and λ_s became larger. From the results, we can see that the proposed variation similarities are useful for the DNN to generate more natural and intelligible speech. In all the following experiments, we fixed $\lambda_m, \lambda_t, \lambda_s = \{1, 5, 5\}$ which demonstrated a good performance, and used $\{1, 0, 0\}$, $\{1, 5, 0\}$, $\{1, 0, 5\}$, and $\{0, 5, 5\}$ to isolate the performance improvement achieved by each contribution.

Table 2: Results of PESQ scores, STOI and SDR values of various algorithms in matched noise conditions.

	SNR (dB)					-5			0			5			10		
	Ob. function	λ_m	λ_t	λ_s	Post-filter	White	Res.	Con.	White	Res.	Con.	White	Res.	Con.	White	Res.	Con.
PESQ	*unprocessed*					1.24	1.32	1.28	1.53	1.64	1.63	1.87	2.04	2.08	2.23	2.42	2.39
	$MMSE-LSA$					1.56	1.41	1.33	2.00	1.80	1.77	2.41	2.26	2.25	2.73	2.61	2.56
	C_{mse}	-	-	-	-	1.91	1.70	2.06	2.41	2.14	2.48	2.76	2.55	2.84	3.03	2.88	3.04
	C_{mse}	-	-	-	GV	1.91	1.69	2.07	2.45	2.19	2.54	2.85	2.64	2.95	3.15	3.01	3.15
	C	1	0	0	GV	2.03	1.77	2.15	2.55	2.31	2.64	2.91	2.74	3.04	3.22	3.11	3.25
	C	1	5	0	GV	2.13	1.91	2.29	2.58	2.40	2.72	2.93	2.81	3.07	3.23	3.14	3.28
	C	1	0	5	GV	2.11	1.88	2.22	2.62	2.43	2.71	2.97	2.84	3.15	3.31	3.22	3.34
	C	1	5	5	GV	**2.18**	**2.02**	**2.38**	**2.65**	**2.52**	**2.82**	**3.00**	**2.92**	**3.16**	**3.32**	**3.24**	**3.39**
	C	0	5	5	GV	1.99	1.82	2.15	2.36	2.26	2.52	2.66	2.58	2.81	2.92	2.86	3.01
STOI	*unprocessed*					0.58	0.51	0.56	0.71	0.64	0.67	0.82	0.75	0.78	0.91	0.85	0.85
	$MMSE-LSA$					0.58	0.45	0.52	0.68	0.58	0.61	0.78	0.71	0.73	0.85	0.80	0.81
	C_{mse}	-	-	-	-	0.66	0.61	0.69	0.78	0.74	0.81	0.86	0.83	0.86	0.91	0.88	0.90
	C_{mse}	-	-	-	GV	0.66	0.61	0.69	0.79	0.75	0.82	0.87	0.84	0.87	0.92	0.89	0.91
	C	1	0	0	GV	0.68	0.62	0.70	0.80	0.77	0.83	0.88	0.85	0.89	0.93	0.91	0.92
	C	1	5	0	GV	**0.75**	**0.71**	0.76	0.83	0.81	0.86	**0.90**	0.87	0.90	**0.94**	**0.92**	0.93
	C	1	0	5	GV	0.71	0.63	0.72	0.82	0.78	0.84	0.89	0.86	0.90	**0.94**	**0.92**	0.93
	C	1	5	5	GV	**0.75**	**0.71**	**0.77**	**0.84**	**0.82**	**0.87**	**0.90**	**0.88**	**0.91**	**0.94**	**0.92**	**0.94**
	C	0	5	5	GV	**0.75**	**0.71**	**0.77**	0.83	0.81	0.85	0.89	0.87	0.90	0.93	0.91	0.92
SDR	$MMSE-LSA$					5.25	-0.53	-1.94	8.53	3.65	2.05	11.13	8.14	6.80	13.31	11.31	10.97
	C_{mse}	-	-	-	-	6.06	2.06	6.44	8.99	5.68	8.59	11.02	8.65	10.65	12.27	10.76	11.97
	C_{mse}	-	-	-	GV	6.61	2.64	7.23	9.94	6.52	9.82	12.58	9.91	12.45	14.66	12.70	14.44
	C	1	0	0	GV	7.21	3.40	7.86	10.51	7.17	10.64	13.40	10.74	13.51	15.88	13.94	15.70
	C	1	5	0	GV	6.83	3.46	7.58	9.88	7.09	10.25	12.60	10.45	13.03	15.01	13.45	15.13
	C	1	0	5	GV	7.52	4.07	8.33	**10.82**	7.91	11.16	**13.81**	11.31	14.20	**16.77**	14.75	16.70
	C	1	5	5	GV	**7.68**	**4.79**	**8.57**	10.75	**8.34**	**11.32**	13.60	**11.53**	**14.21**	16.43	**14.77**	**16.71**
	C	0	5	5	GV	5.10	2.13	5.57	7.57	5.55	7.91	9.60	8.10	9.77	11.16	10.19	11.14

4.2. Performance Evaluation in Matched Noise Conditions

In this experiment, the performances of the MMSE-log spectral amplitude (LSA) estimation algorithm [3] and various configurations of the DNN were compared in the matched noise conditions. Table II shows the PESQ scores, STOI values and SDR values obtained in matched noise conditions. From the results, the DNN-based algorithms outperformed conventional MMSE-LSA algorithm especially in non-stationary noise environments. Also, it is shown that employing the Mel-scale weighted mean square error improved both the predicted perceptual quality and intelligibility of the enhanced speech. This result demonstrates that adopting perceptually motivated non-linear frequency scale to the objective function improves the quality of the enhanced speech.

Moreover, incorporating the variation similarities into the DNN training objective function further improved the performance in terms of both PESQ score and STOI value. In the case of PESQ score, the performance of the DNN was improved with the use of ρ_{temp} and ρ_{spec}. On the other hand, it turned out that ρ_{temp} played more important role to improve the STOI values particularly in low SNR conditions than ρ_{spec}. These results were consistent with the previous studies which reported that the temporal variation is more important in speech intelligibility.

In contrast, the SDR values were improved by adopting ρ_{spec} while ρ_{temp} decreased them. Since the one-third octave bands do not discriminate the frequency bins in each band, detailed spectral representation might be muffled by applying ρ_{temp}. However, when both variation similarities were adopted, the SDR performances of the enhanced speech were also improved compared to those with simple mean square error-based objective function.

Finally, the objective function without C_{wmse} showed worse performance than the conventional objective function C_{mse}. This result shows that while matching temporal and spectral variation similarities improves the enhanced speech quality, the mean square error between enhanced and reference speech time-frequency bins still has crucial role to train the model.

4.3. Performance Evaluation in Mismatched Noise Conditions

In this experiment, the performances of various algorithms were compared in the mismatched noise conditions. Table III shows the PESQ scores, STOI values, and SDR values obtained in various mismatched noise conditions. From the results, we can see that the DNN-based algorithm were more robust to the distortion from realistic noise environments such as Cafeteria and Kids compared to the MMSE-LSA algorithm. In these conditions, the MMSE-LSA algorithm did not separated speech component from the input signal correctly.

The amount of improvement in both the quality and intelligibility were less than that achieved in the matched noise condition. However, the DNN-based speech enhancement algorithm with the proposed objective function still outperformed the conventional algorithm in unseen noise conditions. The PESQ, STOI, and SDR performances of the enhanced speech were improved by employing the Mel-scale weighted mean square error and variation similarities.

It is interesting to see that the incorporation of the spectral variation similarity ρ_{spec} slightly decreased the enhancement performance in the Kids noise environment. This may be due to the characteristics of the Kids noise which has similar spectral shape to the target speech. Since ρ_{spec} emphasizes the spectral peaks and valleys, it made the speech-like noise slightly more noticeable after speech enhancement. In terms of PESQ scores, applying the Mel-scale weight contributed the largest performance improvement while temporal and spectral variation similarity showed equal contribution.

As shown in Table III, the STOI values were enhanced by incorporating ρ_{temp} to the objective function while other techniques did not show any significant effects on predicted intelligibility score. This result once again confirms that the temporal variation of the enhanced speech is more crucial than the spectral variation in terms of the speech intelligibility. The human intelligibility prediction was significantly improved by adopting the temporal variation similarity. Among the rest two sub-costs, Mel-scale weight contributed more than the spectral variation similarity.

Similar to the results in matched noise conditions, the SDR values were increased by adopting C_{wmse} and ρ_{spec} while they were decreased by adding ρ_{temp}. Note that the SDR values of proposed algorithm with $\lambda_m, \lambda_t, \lambda_s = \{0, 5, 5\}$ showed worse SDR performance than conventional algorithm in high-SNR conditions. The results shows that the objective function which combines all the three sub-costs showed best performance with various objective measures.

Figure 3 shows the spectrograms of an utterance enhanced by conventional and proposed DNN-based algorithms. From this figure, it is shown that the proposed algorithm effectively reduced the noise from the original speech while the speech distortion was minimized. We also uploaded several audio files in (https://mspl.gist.ac.kr/wp-content/uploads/2017/10/Demo.zip).

Table 3: Results of PESQ scores, STOI and SDR values of various algorithms in mismatched noise conditions.

	SNR (dB)					-5			0			5			10		
	Ob. function	λ_m	λ_t	λ_s	Post-filter	Cafe.	Kids	Str.	Cafe.	Kids	Str.	Cafe.	Kids	Str.	Cafe.	Kids	Str.
PESQ	*unprocessed*					1.43	1.27	1.66	1.75	1.71	1.99	2.13	2.11	2.25	2.50	2.38	2.63
	$MMSE-LSA$					1.43	1.09	1.95	1.86	1.54	2.29	2.28	1.98	2.53	2.62	2.36	2.83
	C_{mse}	-	-	-	-	1.61	1.66	1.86	2.04	1.98	2.35	2.50	2.46	2.63	2.86	2.68	2.96
	C_{mse}	-	-	-	*GV*	1.63	1.68	1.87	2.09	2.06	2.42	2.59	2.55	2.74	2.97	2.79	3.08
	C	1	0	0	*GV*	1.69	1.74	1.95	2.21	2.11	2.51	2.67	**2.62**	2.83	3.06	**2.86**	3.18
	C	1	5	0	*GV*	1.72	**1.78**	2.04	2.24	**2.14**	2.56	2.66	2.59	2.87	3.03	**2.86**	3.20
	C	1	0	5	*GV*	1.74	1.72	1.91	2.28	2.08	2.55	2.73	2.56	2.89	**3.12**	2.85	**3.26**
	C	1	5	5	*GV*	**1.80**	1.73	**2.09**	**2.31**	2.11	**2.62**	**2.74**	2.54	**2.91**	3.10	2.84	**3.26**
	C	0	5	5	*GV*	1.74	1.66	2.04	2.12	1.96	2.38	2.49	2.32	2.62	2.79	2.58	2.91
STOI	*unprocessed*					0.54	0.57	0.66	0.65	0.70	0.76	0.77	0.83	0.83	0.85	0.89	0.90
	$MMSE-LSA$					0.48	0.48	0.63	0.59	0.62	0.72	0.72	0.75	0.79	0.80	0.82	0.86
	C_{mse}	-	-	-	-	0.58	0.65	0.65	0.72	0.77	0.80	0.82	0.86	0.85	0.88	0.90	0.91
	C_{mse}	-	-	-	*GV*	0.58	0.65	0.66	0.73	0.78	0.81	0.83	0.88	0.86	0.89	0.91	0.92
	C	1	0	0	*GV*	0.59	0.66	0.67	0.74	0.79	0.82	0.84	**0.89**	0.88	0.90	**0.93**	0.93
	C	1	5	0	*GV*	0.63	**0.67**	0.74	0.77	**0.80**	0.85	0.86	**0.89**	0.89	0.91	**0.93**	**0.94**
	C	1	0	5	*GV*	0.60	0.65	0.68	0.76	0.79	0.83	0.86	**0.89**	0.89	0.91	**0.93**	**0.94**
	C	1	5	5	*GV*	0.64	0.66	0.74	**0.78**	0.79	**0.86**	**0.87**	**0.89**	**0.90**	**0.92**	**0.93**	**0.94**
	C	0	5	5	*GV*	**0.65**	0.66	**0.76**	0.77	0.78	0.85	0.85	0.87	0.89	0.9	0.91	0.93
SDR	$MMSE-LSA$					-1.24	-4.85	1.80	3.91	0.13	7.59	8.16	4.89	9.90	11.49	9.64	13.11
	C_{mse}	-	-	-	-	0.31	0.77	4.99	4.79	4.95	7.87	8.05	8.40	10.07	10.55	10.47	11.65
	C_{mse}	-	-	-	GV	0.96	1.12	5.78	5.59	5.77	9.15	9.35	9.69	11.76	12.50	12.60	14.16
	C	1	0	0	*GV*	1.71	**1.48**	6.39	6.39	**6.19**	10.06	10.14	**10.53**	12.77	13.74	**13.93**	15.59
	C	1	5	0	*GV*	1.05	0.86	6.27	5.76	5.49	9.60	9.50	9.68	12.34	12.95	13.29	15.08
	C	1	0	5	*GV*	**2.97**	1.17	6.74	**7.13**	5.83	**10.63**	**10.72**	10.05	**13.46**	**14.43**	13.89	**16.86**
	C	1	5	5	*GV*	2.14	0.64	**7.27**	6.68	5.27	10.55	10.50	9.63	**13.46**	14.17	13.89	16.68
	C	0	5	5	*GV*	-0.47	-0.53	4.04	3.69	3.3	6.86	6.94	6.69	9.14	9.57	9.33	10.94

4.4. Subjective Test Results

Additionally, we performed a subjective listening test to compare the performance of the proposed techniques with the conventional objective function. Ten listeners participated and were presented with 45 randomly selected sentences

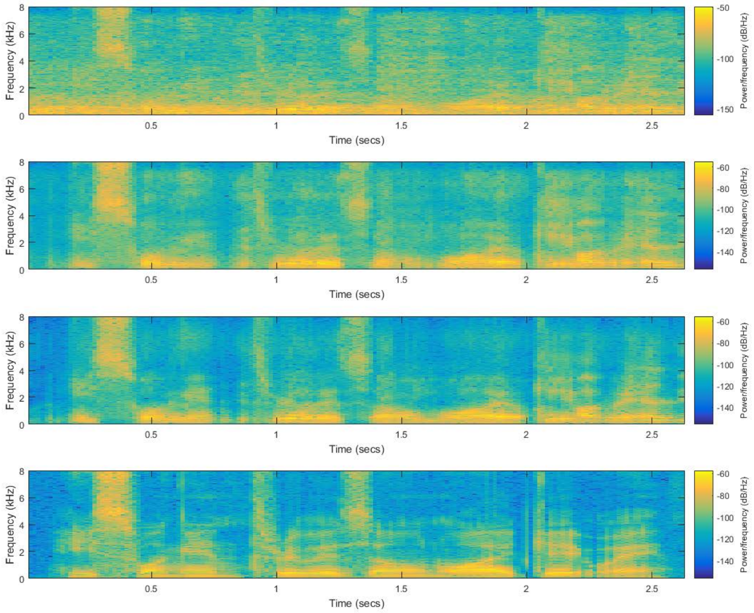

Figure 3: The spectrogram of original speech with cafeteria noise (-5 dB SNR), enhanced speech by conventional DNN-based algorithm, enhanced speech by the proposed algorithm, and the corresponding clean speech, respectively.

in the SNR range of {-5, 0, 5} dB corrupted by the {cafeteria, kids, street} noises. In the test, each listener was provided with speech samples enhanced by the network with the conventional objective function and the proposed objective function with $\lambda_m, \lambda_t, \lambda_s = \{1, 5, 5\}$. Listeners were asked to choose the preferred one for each pair of speech samples in terms of perceptual speech quality. Two samples in each pair were given in arbitrary order.

The results are shown in Figure 4. It can be seen that the quality of the speech enhanced by the proposed algorithm was better than that using the conventional algorithm in all SNR values. These results imply that the proposed algorithm enhances not only the objective measures but also the perceived quality.

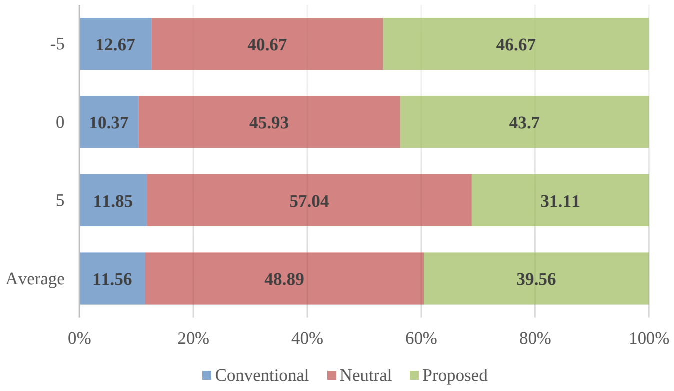

Figure 4: Results of preference test (%) comparing the speech quality for the conventional and proposed algorithms with various SNR values.

5. Conclusions

In this paper, we have proposed a novel objective function for DNN-based speech enhancement to equalize the temporal and spectral variations of the enhanced speech. The proposed algorithm incorporates a perceptually motivated non-linear frequency weight and variation similarities between the enhanced and clean speech spectral trajectories. From the experimental results, it has been found that the proposed algorithm outperformed the conventional DNN-based speech enhancement algorithm in terms of the objective measures as well as the subjective listening quality.

The future work will focus on employing novel model structures and training techniques. The recent studies show that the performance of the deep learning models could be further improved by the proper training scheme [46, 47, 48]. Also, as in speech recognition, the sequence-to-sequence model such as the LSTM or GRU may be better to describe the speech characteristics. Finally, the spatial information of the target and background noise will be useful to improve the speech quality in the interfering speaker environments such as the Kids noise.

Acknowledgement

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2015R1A2A1A15054343).

References

[1] P. C. Loizou, Speech Enhancement: Theory and Practice, Boca Raton, FL, USA: CRC, 2013.

[2] N. S. Kim, J. H. Chang, Statistical model based techniques for robust speech communication, in: Recent Advances in Robust Speech Recognition Technology, 2010, pp. 114–130.

[3] Y. Ephraim, D. Malah, Speech enhancement using minimum mean square log spectral amplitude estimator, IEEE Trans. Acoust., Speech, Signal Process. ASSP-33 (2) (1985) 443–445.

[4] N. S. Kim, J. H. Chang, Spectral enhancement based on global soft decision, IEEE Signal Process. Lett. 7 (5) (2000) 108–110.

[5] I. Cohen, B. Berdugo, Speech enhancement for non-stationary noise environments, Signal Process. 81 (11) (2001) 2403–2418.

[6] I. Cohen, Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging, IEEE Trans. Speech, Audio Process. 11 (5) (2003) 465–475.

[7] J.-M. Kum, Y.-S. Park, J.-H. Chang, Improved minima controlled recursive averaging technique using conditional maximum a posteriori criterion for speech enhancement, Digital Signal Process. 20 (6) (2012) 1572–1578.

[8] J.-H. Chang, Noisy speech enhancement based on improved minimum statistics incorporating acoustic environment-awareness, Digital Signal Process. 23 (4) (2013) 1233–1238.

[9] A. Mohamed, G. E. Dahl, G. Hinton, Acoustic modeling using deep belief networks, IEEE Trans. Audio, Speech, Language Process. 20 (1) (2012) 14–22.

[10] R. Salakhutdinov, G. Hinton, Using deep belief nets to learn covariance kernels for gaussian processes, in: Advances in Neural Inform. Process, Vol. 20, 2007, pp. 1–8.

[11] F. Seide, G. Li, D. Yu, Conversational speech transcription using context-depdendent deep neural networks, in: Proc. Interspeech, 2011, pp. 437–440.

[12] X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection, IEEE Trans. Audio, Speech, Language Process. 21 (4) (2013) 697–710.

[13] T. G. Kang, K. Kwon, J. W. Shin, N. S. Kim, NMF-based target source separation using deep neural network, IEEE Signal Process. Lett. 22 (2) (2015) 229–233.

[14] Y. Wang, D. Wang, Towards scaling up classification-based speech separation, IEEE/ACM Trans. Audio, Speech, Language Process. 21 (7) (2013) 1381–1390.

[15] X. Lu, Y. Tsao, S. Matsuda, C. Hori, Speech enhancement based on deep denoising autoencoder, in: Proc. Interspeech, 2013, pp. 436–440.

[16] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, An experimental study on speech enhancement based on deep neural networks, IEEE Signal Process. Lett. 21 (1) (2014) 65–68.

[17] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, A regression approach to speech enhancement based on deep neural networks, IEEE/ACM Trans. Audio, Speech, Language Process. 23 (1) (2015) 7–19.

[18] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smargadis, Joint optimization of masks and deep recurrent neural networks for monaural source

separation, IEEE/ACM Trans. Audio, Speech, Language Process. 23 (12) (2015) 2136–2147.

[19] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, T. Zhang, Learning spectral mapping for speech dereverberation and denoising, IEEE/ACM Trans. Audio, Speech, Language Process. 23 (6) (2015) 982–992.

[20] Y. Wang, A. Narayanan, D. L. Wang, On training targets for supervised speech separation, IEEE/ACM Trans. Audio, Speech, Language Process. 22 (12) (2014) 1849–1858.

[21] H. Erdogan, J. R. Hershey, S. Watanabe, J. L. Roux, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, in: Proc. ICASSP, 2015, pp. 708–712.

[22] D. S. Williamson, Y. Wang, D. L. Wang, Complex ratio masking for monaural speech separation, IEEE/ACM Trans. Audio, Speech, Language Process. 24 (3) (2016) 483–492.

[23] X.-L. Zhang, D. L. Wang, A deep ensemble learning method for monaural speech separation, IEEE/ACM Trans. Audio, Speech, Language Process. To be appeared.

[24] Y. Xu, J. Du, Z. Huang, L.-R. Dai, C.-H. Lee, Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement, in: Proc. Interspeech, 2015, pp. 1508–1512.

[25] J. Du, Y. Tu, L.-R. dai, C.-H. Lee, A regression approach to single-channel speech separation via high-resolution deep neural networks, IEEE/ACM Trans. Audio, Speech, Lang. Process. 24 (8).

[26] J. Du, Y. Xu, Hierarchical deep neural network for multivariate regression, Pattern Recognition 63 (2017) 149–157.

[27] R. L. Goldsworthy, J. E. Greenberg, Analysis of speech-based speech transmission index methods with implications for nonlinear operations, J. Acoust. Soc. Amer. 116 (6) (2004) 3679–3689.

[28] J. M. Kates, K. H. Arehart, Coherence and the speech intelligibility index, J. Acoust. Soc. Amer. 117 (4) (2005) 2224–2237.

[29] B. Xia, C. Bao, Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification, Speech Commun. 60 (2014) 13–29.

[30] P. G. Shivakumar, P. Georgiou, Perception optimized deep denoising autoencoders for speech enhancement, in: Proc. Interspeech, 2016, pp. 3743–3747.

[31] H. S. Kim, Y. M. Cho, H.-J. Kim, Speech enhancement via Mel-scale Wiener filtering with a frequency-wise voice activity detector, J. Mech. Sci. Technology 21 (5) (2007) 708–722.

[32] C.H.Taal, R.C.Hendriks, R.Heusdens, An algorithm for intelligibility prediction of time-frequency weighted noisy speech, IEEE Trans. Audio, Speech, Language Process. 21 (7) (2011) 2125–2136.

[33] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, Vol. 385 of Studies in Computational Intelligence, Springer, 2012.

[34] J. Chung, Ç. Gülçehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, Tech. Rep. Arxiv report 1412.3555, Université de Montréal, presented at the Deep Learning workshop at NIPS2014 (2014).

[35] P. Voigtlaender, P. Doetsch, S. Wiesler, R. Schluter, H. Ney, Sequence-discriminative training of recurrent neural networks, in: Proc. ICASSP, 2015, pp. 2100–2104.

[36] L. Sun, J. Du, L. R. Dai, C. H. Lee, Multiple-target deep learning for lstm-rnn based speech enhancement, in: Proc. HSCMA, 2017, pp. 136–140.

[37] M. Delfarah, D. L. Wang, Features for masking-based monaural speech separation in reverberant conditions, IEEE/ACM Trans. Audio, Speech, Language Process. 25 (5) (2017) 1085–1094.

[38] A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: ICML, 2013.

[39] Y. Bengio, Practical recommendation for gradient-based training of deep architectures, in: G. Montavon, G. Orr, K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade.

[40] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Pearson Education, 2013.

[41] ITU, Test signals for use in telephonometry ITU-T Rec. P. 501,, 2012.

[42] A. Varga, H. J. M. Steeneken, Assessment for automatic speech recognition: Ii.noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Commun. 12 (3) (1993) 247–251.

[43] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, Theano: A cpu and gpu math complier in python, in: Scientific Comput. with Python Conf. (SciPy), 2010, pp. 3–9.

[44] ITU, Perceptual evalulation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Rec. P. 862 (2000).

[45] E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation, IEEE Trans. Audio, Speech, and Language Process. 14 (4) (2006) 1462–1469.

[46] Y. Dauphin, R. Pascanu, Ç. Gülçehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, arXiv abs/1406.2572.

[47] L. J. Ba, B. J. Frey, Adaptive dropout for training deep neural networks, in: Advances in Neural Information Processing Systems, 2013, pp. 3084–3092.

[48] L. Wan, M. D. Zeiler, S. Zhang, Y. LeCun, R. Fergus, Regularization of neural networks using dropconnect, in: Int. Conf. Mach. Learning,, 2013, pp. 1058–1066.

Appendix

In this Appendix we present the detail on deriving $\frac{\partial C}{\partial \mathbf{x}_t(f)}$. The gradient of the proposed objective function is given by the sum of the separate gradients of the three sub-costs as follows:

$$\frac{\partial C}{\partial \mathbf{x}_t(f)} = \frac{\partial C_{wmse}}{\partial \mathbf{x}_t(f)} - \lambda_t \frac{\partial \rho_{temp}}{\partial \mathbf{x}_t(f)} - \lambda_s \frac{\partial \rho_{spec}}{\partial \mathbf{x}_t(f)}. \tag{28}$$

Similar to the conventional mean square error, $\frac{\partial C_{wmse}}{\partial \mathbf{x}_t(f)}$ is given by

$$\frac{\partial C_{wmse}}{\partial \mathbf{x}_t(f)} = \frac{2}{T} w(f)(\mathbf{x}_t(f) - \mathbf{y}_t(f)). \tag{29}$$

The gradients of the second and third sub-costs are given by (30)–(32) and (33)–(35), respectively.

After $\{\frac{\partial C}{\partial \mathbf{x}_t(f)}\}$ are derived, the gradient of the proposed objective function with respect to each network parameter θ is obtained as

$$\frac{\partial C}{\partial \theta} = \sum_{t=1}^{T} \sum_{f=1}^{F} \frac{\partial C}{\partial \mathbf{x}_t(f)} \frac{\partial \mathbf{x}_t(f)}{\partial \theta} \tag{36}$$

and the usual back-propagation algorithm is applied.

$$
\begin{aligned}
\frac{\partial \rho_{temp}}{\partial \mathbf{x}_t(f)} &= \frac{1}{(T-N+1)H} \sum_{\tau=1}^{T-N+1} \sum_{h=1}^{H} \frac{\partial \rho_{temp}(\tau, h)}{\partial \mathbf{x}_t(f)} \\
&= \frac{1}{(T-N+1)H} \sum_{\tau=1}^{T-N+1} \sum_{h=1}^{H} \frac{\partial \rho_{temp}(\tau, h)}{\partial \tilde{\mathbf{x}}_t(h)} \frac{\partial \tilde{\mathbf{x}}_t(h)}{\partial \mathbf{x}_t(f)} \qquad (30)
\end{aligned}
$$

$$
\frac{\partial \rho_{temp}(\tau, h)}{\partial \tilde{\mathbf{x}}_t(h)} = \begin{cases} \frac{\tilde{\mathbf{y}}_t(h) - \mu_{\breve{\mathbf{y}}_{\tau,h}}}{||\breve{\mathbf{x}}_{\tau,h} - \mu_{\breve{\mathbf{x}}_{\tau,h}} \mathbf{1}_N||\ ||\breve{\mathbf{y}}_{\tau,h} - \mu_{\breve{\mathbf{y}}_{\tau,h}} \mathbf{1}_N||} - \rho_{temp}(\tau, h) \frac{\tilde{\mathbf{x}}_t(h) - \mu_{\breve{\mathbf{x}}_{\tau,h}}}{||\breve{\mathbf{x}}_{\tau,h} - \mu_{\breve{\mathbf{x}}_{\tau,h}} \mathbf{1}_N||^2}, & \text{if } \tilde{\mathbf{x}}_t(h) \in \breve{\mathbf{x}}_{\tau,h}, \\ 0, & \text{otherwise.} \end{cases} \qquad (31)
$$

$$
\frac{\partial \tilde{\mathbf{x}}_t(h)}{\partial \mathbf{x}_t(f)} = \begin{cases} \frac{1}{2\tilde{\mathbf{x}}_t(h)}\ \mathbf{s}(f)\ exp(\mathbf{x}_t(f)\mathbf{s}(f) + \mathbf{m}(f)), & B(h, f) = 1, \\ 0, & B(h, f) = 0. \end{cases} \qquad (32)
$$

$$
\frac{\partial \rho_{spec}}{\partial \mathbf{x}_t(f)} = \frac{1}{T} \frac{\partial \rho_{spec}(t)}{\partial \tilde{\mathbf{x}}'_t(f)} \frac{\partial \tilde{\mathbf{x}}'_t(f)}{\partial \mathbf{x}_t(f)}, \qquad (33)
$$

$$
\frac{\partial \rho_{spec}(t)}{\partial \tilde{\mathbf{x}}'_t(f)} = \frac{\tilde{\mathbf{y}}'_t(f) - \mu_{\tilde{\mathbf{y}}'_t}}{||\tilde{\mathbf{x}}'_t - \mu_{\tilde{\mathbf{x}}'_t} \mathbf{1}_F||\ ||\tilde{\mathbf{y}}'_t - \mu_{\tilde{\mathbf{y}}'_t} \mathbf{1}_F||} - \rho_{spec}(t) \frac{\tilde{\mathbf{x}}'_t(f) - \mu_{\tilde{\mathbf{x}}'_t}}{||\tilde{\mathbf{x}}'_t - \mu_{\tilde{\mathbf{x}}'_t} \mathbf{1}_F||^2}, \qquad (34)
$$

$$
\frac{\partial \tilde{\mathbf{x}}'_t(f)}{\partial \mathbf{x}_t(f)} = \frac{1}{2\tilde{\mathbf{x}}'_t(f)}\ \mathbf{s}(f)\ exp(\mathbf{x}_t(f)\mathbf{s}(f) + \mathbf{m}(f)). \qquad (35)
$$