Interspeech 2022
18-22 September 2022, Incheon, Korea

iDeepMMSE: An Improved Deep Learning Approach to MMSE Speech and Noise Power Spectrum Estimation for Speech Enhancement

Minseung Kim, Hyungchan Song, Sein Cheong and Jong Won Shin

School of Electrical Engineering and Computer Science,
Gwangju Institute of Science and Technology, Gwangju 61005, Korea

kms0603@gist.ac.kr, shchan420@gm.gist.ac.kr, seiinjung@gm.gist.ac.kr, jwshin@gist.ac.kr

Abstract

Deep learning approaches have been successfully applied to single channel speech enhancement exhibiting significant performance improvement. Recently, approaches unifying deep learning techniques into a statistical speech enhancement framework were proposed, including Deep Xi and DeepMMSE in which *a priori* signal-to-noise ratios (SNRs) were estimated by deep neural networks (DNNs) and noise power spectral density (PSD) and spectral gain functions were computed with estimated parameters. In this paper, we propose an improved DeepMMSE (iDeepMMSE) which estimates the speech PSD and speech presence probability as well as the *a priori* SNR using a DNN for MMSE estimation of the speech and noise PSDs. The *a priori* and *a posteriori* SNRs are refined with the estimated PSDs, which in turn are used to compute spectral gain function. We also replaced the DNN architecture with the Conformer which efficiently captures the local and global sequential information. Experimental results on the Voice Bank-DEMAND dataset and Deep Xi dataset showed the proposed iDeepMMSE outperformed the DeepMMSE in terms of the perceptual evaluation of speech quality (PESQ) scores and composite objective measures.

Index Terms: Speech enhancement, Conformer, Deep Xi, DeepMMSE, MMSE estimation

1. Introduction

In many speech applications such as voice communication [1], speech recognition [2], speaker verification [3], and hearing aid [4], noisy speech often impairs user satisfaction. Speech enhancement aims to improve the perceived quality and intelligibility of noisy speech by reducing the effects of background noises while minimizing speech distortion.

Many speech enhancement techniques have been proposed for the last decades [5–22]. Among them, statistical model-based speech enhancement techniques [5–12] which employ clean speech estimators derived from various optimisation criteria including a Wiener filter, minimum mean square error short-time spectral amplitude (MMSE-STSA) [5], and minimum mean square error log-spectral amplitude (MMSE-LSA) [6], exhibited decent performances. These minimum mean-square error (MMSE) estimators often require *a priori* signal-to-noise ratio (SNR) and *a posteriori* SNR estimates. In recent years, deep learning approaches have been widely studied for speech enhancement and shown significant performance improvements. One of the popular approaches to deep learning-based speech enhancement is to estimate target masks from the noisy speech representation and multiply them to noisy spectra directly. Various target masks [15] have been investigated including ideal binary mask (IBM), ideal ratio mask (IRM), ideal amplitude mask (IAM), phase sensitive mask (PSM) [16], and complex ideal ratio mask (cIRM) [17].

Recently, several studies have proposed to incorporate deep learning techniques into statistical speech enhancement frameworks. In [18], *a priori* SNR is estimated by a deep learning network instead of target masks and is used to compute spectral gains based on the statistical framework. Since the instantaneous *a priori* SNR has a large dynamic range, this method, which is called Deep Xi, adopted a compression function to facilitate training. Various gain functions and deep neural network (DNN) architectures have been applied to the Deep Xi framework in the following studies [19, 20]. The DeepMMSE [21] employs the Deep Xi framework to MMSE noise power spectral density (PSD) estimator. While the DeepMMSE showed excellent noise PSD tracking performance, the *a priori* SNR estimate that they have adopted for speech enhancement was the maximum likelihood (ML) estimate which relies only on the noise PSD estimates and is irrelevant to the speech PSD estimate, conveying essentially the same information with the *a posteriori* SNR.

In this paper, we propose an improved DeepMMSE (iDeepMMSE) which exploits the DNN to estimate the speech PSD and speech presence probability (SPP) on top of the *a priori* SNR. The MMSE speech and noise PSD estimators incorporating speech presence uncertainty are employed, in which the parameters are estimated by a DNN. For more sophisticated temporal sequence modelling, we adopted the convolution-augmented transformer (Conformer) [23] as the DNN architecture. With the speech and noise PSD estimates computed with DNN-based estimated statistics, we refine the *a priori* SNR and *a posteriori* SNR to evaluate the final gain function. The proposed iDeepMMSE exhibited superior speech enhancement performance compared to the DeepMMSE in our experiments on the Voice Bank-DEMAND dataset and Deep Xi dataset.

2. Deep Xi and DeepMMSE Approaches for Speech Enhancement

2.1. Signal model

Assuming that the speech $S(l,k)$ is corrupted by additive noise $N(l,k)$, the observed noisy speech $Y(l,k)$ in the short-time Fourier transform (STFT) domain is given as

$$Y(l,k) = S(l,k) + N(l,k), \tag{1}$$

where $Y(l,k)$, $S(l,k)$, and $N(l,k)$ are the STFT coefficients for $y(t)$, $s(t)$, and $n(t)$ at frame l and frequency k, respectively. We assume that the STFT coefficients for speech and noise follow zero-mean complex Gaussian distribution and are uncorrelated. Then, the PSDs for noisy speech, clean and noise,

10.21437/Interspeech.2022-964

$\Phi_y(l,k)$, $\Phi_s(l,k)$ and $\Phi_n(l,k)$, can be related as $\Phi_y(l,k) = \Phi_s(l,k) + \Phi_n(l,k)$.

The MMSE-LSA clean speech estimator which minimizes the mean square error between the clean and estimated speech log-magnitude spectra can be expressed as

$$\left|\widehat{S}(l,k)\right| = G(l,k) \cdot |Y(l,k)|, \tag{2}$$

in which $G(l,k)$ is the gain function given as [6]

$$G(l,k) = \frac{\xi(l,k)}{\xi(l,k)+1} \exp\left\{\frac{1}{2}\int_{v(l,k)}^{\infty} \frac{e^{-t}}{t} dt\right\}, \tag{3}$$

where $v(l,k) = [\xi(l,k)/(\xi(l,k)+1)]\gamma(l,k)$, $\xi(l,k)$ is the *a priori* SNR and $\gamma(l,k)$ is the *a posteriori* SNR, which are defined by

$$\xi(l,k) = \frac{\Phi_s(l,k)}{\Phi_n(l,k)}, \tag{4}$$

$$\gamma(l,k) = \frac{|Y(l,k)|^2}{\Phi_n(l,k)}. \tag{5}$$

2.2. Deep Xi

To evaluate the gain function in (3), the *a priori* and *a posteriori* SNRs should be estimated. Deep Xi [18–20] exploits DNNs to estimate the *a priori* SNRs from the input magnitude spectrogram $|Y|$ and then use the estimates to evaluate the gain function. The training target for the *a priori* SNR is set to be the instantaneous estimate given by

$$\xi^{inst} = \frac{|S(l,k)|^2}{|N(l,k)|^2}. \tag{6}$$

Since the instantaneous *a priori* SNR has a large dynamic range, it is difficult to train a DNN to estimate the instantaneous *a priori* SNR directly [18]. To circumvent this difficulty, the cumulative distribution function (CDF) of $\xi_{dB}^{inst} = 10\log_{10}(\xi^{inst})$ was used as a mapping function to compress the dynamic range. Assuming that ξ_{dB}^{inst} follows a normal distribution, the training target becomes

$$\bar{\xi}(l,k) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\xi_{dB}^{inst}(l,k) - \mu(k)}{\sigma(k)\sqrt{2}}\right)\right], \tag{7}$$

where $\text{erf}(\cdot)$ is the error function, $\bar{\xi}$ is the mapped *a priori* SNR and the parameters $\mu(k)$ and $\sigma(k)$ for each frequency are estimated from the histogram of the spectral components for noisy speech randomly selected from training data.

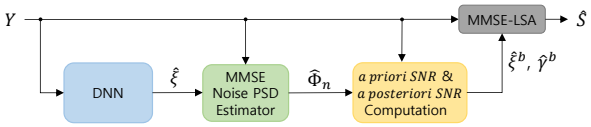
In the inference phase, the mapped *a priori* SNR $\bar{\xi}(l,k)$ is estimated from $|Y|$ using a DNN as

$$\hat{\bar{\xi}} = \text{DNN}(|Y|), \tag{8}$$

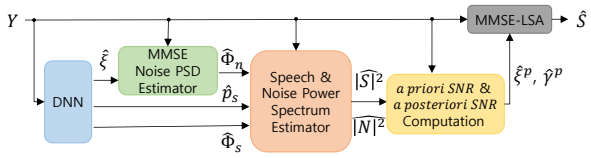
where $\text{DNN}(\cdot)$ can be any DNN architecture such as long short-term memory (LSTM), temporal convolutional network (TCN) and Conformer. The estimate for the *a priori* SNR can be obtained using the inverse function of (7) as

$$\widehat{\xi}(l,k) = 10^{\{\sigma(k)\sqrt{2}\text{erf}^{-1}(2\hat{\bar{\xi}}(l,k)-1)+\mu(k)\}/10}. \tag{9}$$

In Deep Xi [18], the gain function in (3) is evaluated with the DNN-based $\widehat{\xi}(l,k)$, in which the *a posteriori* SNR $\widehat{\gamma}(l,k)$ is computed as $\widehat{\xi}(l,k) + 1$ following the ML estimate of the *a priori* SNR.



(a) *Block diagram of the baseline DeepMMSE framework.*



(b) *Block diagram of the proposed iDeepMMSE framework.*

Figure 1: *Block diagrams of the baseline DeepMMSE and proposed iDeepMMSE frameworks.*

2.3. DeepMMSE

DeepMMSE [21] follows the framework of Deep Xi to estimate noise PSD, and refines the estimates for the *a priori* and *a posteriori* SNRs with the noise PSD estimate. The MMSE noise power spectrum estimator is given as

$$\begin{aligned}\widehat{|N|^2} &= E(|N|^2|Y,\xi,\gamma) \\ &= \left(\frac{1}{(1+\xi)^2} + \frac{\xi}{(1+\xi)\gamma}\right)|Y|^2,\end{aligned} \tag{10}$$

where $E(\cdot)$ denotes the expectation. Using $\widehat{|N|^2}$ in the current frame, the noise PSD is obtained by applying temporal recursive smoothing [10] as

$$\widehat{\Phi}_n(l,k) = \alpha_n \Phi_n(l-1,k) + (1-\alpha_n)\widehat{|N|^2}(l,k), \tag{11}$$

where α_n is a smoothing parameter. Then, the *a priori* and *a posteriori* SNRs are refined by utilizing $\widehat{\Phi}_n(l,k)$

$$\widehat{\gamma}^b(l,k) = \frac{|Y(l,k)|^2}{\widehat{\Phi}_n(l,k)}, \tag{12}$$

$$\widehat{\xi}^b(l,k) = \max(\widehat{\gamma}^b(l,k) - 1, 0), \tag{13}$$

in which the superscript ^{b} denotes the baseline method, and used to evaluate the gain function in (3). It is noted that α_n was set to zero and $\widehat{\gamma}(l,k) = \widehat{\xi}(l,k) + 1$ were used to evaluate (10) for the experiments in [21], which resulted in the same gain function with Deep Xi.

3. Proposed Improved DeepMMSE Method

In this paper, we propose to estimate speech PSD and *a posteriori* SPP as well as *a priori* SNR using a DNN, compute MMSE estimates for speech and noise power spectra, and produce the spectral gains using the *a priori* and *a posteriori* SNRs obtained with speech and noise power spectrum estimates. The overall block diagrams for the baseline DeepMMSE [21] and the proposed iDeepMMSE are shown in Fig. 1. One drawback of the Deep Xi and DeepMMSE may be the *a priori* and *a posteriori* SNRs used to evaluate the gain function convey exactly the same information, although the original definitions of them in (4) imply that the *a priori* information would be more smoothed and stable compared with the *a posteriori* SNR which has information on instantaneous input. Also, the speech presence uncertainty was not incorporated in the Deep Xi and DeepMMSE.

The input and the training target for the *a priori* SNR are the same as those for DeepMMSE. The training target for the *a posteriori* SPP is constructed by thresholding the power of the clean speech used to construct the training data in each time-frequency bin to form binary speech presence masks. Specifically, the time-frequency bins with the power less than -50dB from the maximum level in each utterance are regarded as silence, while other time-frequency bins are considered to be speech active regions as in [8]. As for the training target for speech PSD, dynamic range compression similar to (7) is applied. Specifically, the distribution of $\Phi_s(l,k)$ in the clean speech dataset is modelled as a log-normal distribution, and the CDF of the dB-scale instantaneous speech power spectrum, $\Phi_{s,dB}^{inst}(l,k) = 10\log_{10}(|S(l,k)|^2)$, is used as the mapping function

$$\bar{\Phi}_s(l,k) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\Phi_{s,dB}^{inst}(l,k) - \mu_s(k)}{\sigma_s(k)\sqrt{2}}\right)\right], \quad (14)$$

where the parameters $\mu_s(k)$ and $\sigma_s(k)$ are estimated from the histogram of $|S|^2$ in the speech active time-frequency bins of the clean speech used to construct training data.

In the inference phase of iDeepMMSE, DNN estimates a mapped version of the *a priori* SNR $\bar{\xi}$, a mapped version of the speech PSD $\bar{\Phi}_s$ and *a posteriori* SPP p_s, from the noisy speech magnitude spectrum as

$$[\hat{\bar{\xi}}, \hat{\bar{\Phi}}_s, \widehat{p}_s] = \text{DNN}(|Y|). \quad (15)$$

As in the DeepMMSE, the estimates for *a priori* SNR and speech PSD are obtained by (9) and the inverse function of (14)

$$\widehat{\Phi}_s(l,k) = 10^{\{\sigma_s\sqrt{2}\text{erf}^{-1}(2\hat{\bar{\Phi}}_s(l,k)-1)+\mu_s(k)\}/10}. \quad (16)$$

Let two hypotheses H_0 and H_1 denote speech absence and presence, respectively. The MMSE speech and noise power spectrum estimators under the speech presence uncertainty are given as [7]

$$\begin{aligned}\widehat{|S|^2} =& E(|S|^2|Y,\Phi_s,\Phi_n)\\ =& p(H_0|Y)\cdot E(|S|^2|Y,\Phi_s,\Phi_n,H_0)\\ &+ p(H_1|Y)\cdot E(|S|^2|Y,\Phi_s,\Phi_n,H_1), \quad (17)\\ \widehat{|N|^2} =& E(|N|^2|Y,\Phi_s,\Phi_n)\\ =& p(H_0|Y)\cdot E(|N|^2|Y,\Phi_s,\Phi_n,H_0)\\ &+ p(H_1|Y)\cdot E(|N|^2|Y,\Phi_s,\Phi_n,H_1), \quad (18)\end{aligned}$$

where $p(H_0|Y) = 1 - p(H_1|Y)$ and $p(H_1|Y)$ are *a posteriori* speech absence and presence probability, respectively, and

$$\begin{aligned}E(|S|^2|Y,\Phi_s,\Phi_n,H_0) =& 0, \quad (19)\\ E(|N|^2|Y,\Phi_s,\Phi_n,H_0) =& |Y|^2, \quad (20)\end{aligned}$$

$$\begin{aligned}&E(|S|^2|Y,\Phi_s,\Phi_n,H_1)\\ &= \left(\frac{\Phi_s}{\Phi_n+\Phi_s}\right)^2|Y|^2 + \frac{\Phi_n}{\Phi_s+\Phi_n}\Phi_s, \quad (21)\\ &E(|N|^2|Y,\Phi_s,\Phi_n,H_1)\\ &= \left(\frac{\Phi_n}{\Phi_n+\Phi_s}\right)^2|Y|^2 + \frac{\Phi_s}{\Phi_s+\Phi_n}\Phi_n. \quad (22)\end{aligned}$$

At first, the rough estimate for the noise PSD $\widehat{\Phi}_n$ is obtained via (10)-(11) as in DeepMMSE. With the $\widehat{\Phi}_n$, and the estimates of speech PSD and *a posteriori* SPP obtained by a DNN and an inverse mapping function in (15) and (16), we can evaluate the MMSE speech and noise power spectrum estimators in (17) and (18). Then, the refined speech and noise PSDs can be obtained by applying temporal recursive smoothing to $\widehat{|S|^2}$ and $\widehat{|N|^2}$ as [7],

$$\Phi_s^p(l,k) = \alpha_s\Phi_s^p(l-1,k) + (1-\alpha_s)\widehat{|S|^2}(l,k), \quad (23)$$

$$\Phi_n^p(l,k) = \alpha_n\Phi_n^p(l-1,k) + (1-\alpha_n)\widehat{|N|^2}(l,k), \quad (24)$$

where the superscript ^{p} denotes the proposed method and α_s is a smoothing parameter. Using $\Phi_s^p(l)$ and $\Phi_n^p(l)$, we can refine the estimates for the *a priori* SNR and *a posteriori* SNR as

$$\widehat{\xi}^p(l,k) = \frac{\Phi_s^p(l,k)}{\Phi_n^p(l,k)}, \quad (25)$$

$$\widehat{\gamma}^p(l,k) = \frac{|Y(l,k)|^2}{\Phi_n^p(l,k)}. \quad (26)$$

Finally, the gain function (3) is evaluated using $\widehat{\xi}^p$ and $\widehat{\gamma}^p$.

4. Experiments

4.1. Dataset

To demonstrate the superiority of the proposed method, we used Voice Bank-DEMAND dataset [24] and Deep Xi dataset [25]. In the Voice Bank-DEMAND dataset, the training set consists of 11,572 clean speech recordings from 28 speakers mixed with two artificial and eight real noises. Each clean speech utterance was contaminated by a randomly selected section of one of the noises with a random SNR of 0, 5, 10 or 15 dB. The test set includes 824 clean speech recordings from 2 speakers mixed with one of the 5 types of noise from the DEMAND dataset at 4 SNR levels including 2.5, 7.5, 12.5, and 17.5 dB.

Deep Xi dataset is larger than the Voice Bank-DEMAND dataset. There are 69,708 clean speech recordings and 17,458 noise recordings used to construct the training and validation sets. 1,000 clean speech and noise recordings were randomly selected from the aforementioned sets to construct the validation set, while the rest of the clean speech and noise recordings were used for training. Each clean speech recording was mixed with a randomly selected section of a noise recording at a randomly selected SNR level out of -5 to 20 dB range, with 1 dB increments. For the test set, 10 clean speech utterances were mixed with 4 types of real-world noises at 5 SNR levels, $\{-5, 0, 5, 10, 15\}$ dB, to form a test set of 200 noisy speech signals.

4.2. Experimental setup

As the architecture of the DNN to estimate three statistical variables in (15), the Conformer [23], which can capture both the local and global sequential information by incorporating a depth-wise convolution layer into a Transformer block, was employed. Each Conformer block consists of a multi-head self-attention (MHSA) module and Convolution module sandwiched by two feed-forward modules, followed by layer normalization [26]. After multiple stacks of Conformer blocks, the last block is connected to a fully-connected layer with sigmoidal activation. The number of Conformer blocks was 6 with 4 attention heads, 256

Table 1: *The performance of speech enhancement for the causal and non-causal versions of the DEMUCS [22], DeepMMSE with TCN and Conformer, and proposed iDeepMMSE on the Voice Bank-DEMAND dataset.*

methods	Causal	PESQ	STOI	CSIG	CBAK	COVL
noisy	-	1.97	0.82	3.36	2.44	2.64
DEMUCS [22]	Yes	2.93	**0.95**	4.22	3.25	3.52
DeepMMSE (TCN) [21]	Yes	2.77	0.93	4.14	3.32	3.46
DeepMMSE (Conformer)	Yes	2.94	**0.95**	**4.30**	3.42	3.63
iDeepMMSE (Conformer)	Yes	**3.07**	**0.95**	4.25	**3.55**	**3.67**
DEMUCS [22]	No	3.07	**0.95**	4.31	3.4	3.63
DeepMMSE (TCN) [21]	No	2.95	0.94	4.28	3.46	3.64
DeepMMSE (Conformer)	No	2.97	**0.95**	**4.32**	3.45	3.65
iDeepMMSE (Conformer)	No	**3.09**	**0.95**	4.25	**3.56**	**3.67**

Table 2: *The performance of speech enhancement for the DeepMMSE with TCN structure [20], and the causal and non-causal versions of DeepMMSE with Conformer and proposed iDeepMMSE on the Deep Xi dataset.*

methods	Causal	PESQ	STOI	CSIG	CBAK	COVL
noisy	-	1.24	0.78	2.26	1.81	1.67
DeepMMSE (TCN) [20]	Yes	1.87	0.84	3.16	2.57	2.47
DeepMMSE (Conformer)	Yes	1.99	**0.87**	**3.33**	2.62	2.61
iDeepMMSE (Conformer)	Yes	**2.04**	**0.87**	3.31	**2.70**	**2.63**
DeepMMSE (Conformer)	No	2.01	**0.87**	3.30	2.65	2.60
iDeepMMSE (Conformer)	No	**2.07**	**0.87**	**3.32**	**2.72**	**2.66**

attention dimensions and 256 feed-forward dimensions. The dimension of DNN output was 257 for baseline DeepMMSE and 771 ($= 257 \times 3$) for proposed iDeepMMSE. The other settings for the network structure follow the same as described in [26]. For causal configuration, a 1-D depth-wise causal convolution was applied inside the Convolution module.

The sampling rate was 16 kHz, and the frame size was 32ms with 50% overlap. A square-root Hann window was used for analysis and synthesis, and the 512 point STFT was applied. The smoothing parameters α_s and α_n were set to zero.

For the training stage, the loss function was the binary cross entropy (BCE) for DeepMMSE as in [21], and the weighted sum of BCE losses for three outputs for iDeepMMSE, in which the weights for the *a priori* SNR, the speech PSD, and *a posteriori* SPP were 5:5:1. We used Adam optimizer [27] with a learning rate $1e^{-3}$. Gradients were clipped to $[-1, 1]$ range. The number of epochs was 200 and the mini-batch size was 8. Each example in a mini-batch was elongated to have the same length with the longest utterance in the mini-batch by zero-padding.

4.3. Experimental results

The performance of the speech enhancement was evaluated in terms of the perceptual evaluation of speech quality (PESQ) scores [28], the short-time objective intelligibility (STOI) [29], and the composite objective measures [30] for signal distortion (CSIG), residual noise (CBAK), and overall quality (COVL). Table 1 shows the performances for the causal and non-causal versions of the DEMUCS [22], DeepMMSE with TCN and Conformer architecture, and proposed iDeepMMSE evaluated on the Voice Bank-DEMAND dataset. In the causal setup, the average PESQ score for the proposed method was improved by 0.14, 0.30, and 0.13 compared with those for the DEMUCS and DeepMMSE with TCN and Conformer architectures, respectively. The STOI and COVL for the proposed method were also similar or slightly better. In the non-causal setup, the proposed method slightly outperformed others in terms of PESQ scores and COVL. It is noted that the performance gap between causal and non-causal configurations was small when Conformer was adopted, for both DeepMMSE and iDeepMMSE.

Table 3: *The PESQ scores of DeepMMSE and proposed iDeepMMSE depending on the SNR for the Deep Xi dataset.*

methods	causal	SNR				
		-5	0	5	10	15
noisy	-	1.05	1.07	1.13	1.31	1.64
DeepMMSE (Conformer)	Yes	1.25	1.54	1.93	2.39	2.84
iDeepMMSE (Conformer)	Yes	**1.26**	**1.57**	**1.99**	**2.44**	**2.92**
DeepMMSE (Conformer)	No	1.26	1.54	1.95	2.40	2.88
iDeepMMSE (Conformer)	No	**1.27**	**1.60**	**2.04**	**2.50**	**2.95**

Table 2 presents the experimental results on the Deep Xi dataset for the DeepMMSE with TCN structure reported in [20] and the causal and non-causal versions of the DeepMMSE with Conformer structure and the proposed iDeepMMSE. We can see that the DeepMMSE with Conformer structure outperformed that with TCN, and the proposed iDeepMMSE showed slightly better PESQ scores and COVL in both causal and non-causal setups. Table 3 shows the PESQ scores depending on the SNR for the same experiment. We can see that the proposed iDeepMMSE outperformed conventional DeepMMSE in all SNR conditions, and the performance improvement was more pronounced in high SNRs for both causal and non-causal configurations.

5. Conclusions

In this paper, we propose an improved DeepMMSE in which DNN estimates the speech PSD and *a posteriori* SPP on top of the *a priori SNR*, and the spectral gain function is obtained from the *a priori* and *a posteriori* SNRs computed with the MMSE speech and noise spectrum estimates. Unlike the DeepMMSE in which the *a priori* and *a posteriori* SNRs essentially have the same information, the proposed *a priori* and *a posteriori* SNR estimators may have different information as in the original definition of them. We also employed Conformer architecture which can efficiently capture the local and global sequential information to further improve the performance. Experimental results showed the proposed iDeepMMSE outperformed the DeepMMSE in terms of PESQ scores and composite metrics for the Voice Bank-DEMAND dataset and Deep Xi dataset.

6. Acknowledgements

This work was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2021-0-01835) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) and by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)).

7. References

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Chichester, U.K.: John Wiley & Sons, 2006.

[2] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood-Cliffs, NJ, USA: Prentice-Hall, 1993.

[3] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, Apr. 1976.

[4] J. M. Kates, *Digital Hearing Aids*. San Diego, CA, USA: Plural Publishing, 2008.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[6] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[7] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108–110, May 2000.

[8] M. Krawczyk-Becker and T. Gerkmann, "On speech enhancement under PSD uncertainty," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1144–1153, Jun. 2018.

[9] M. Kim and J. W. Shin, "Improved speech enhancement considering speech PSD uncertainty," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1939–1951, Jun. 2022.

[10] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[11] J. W. Shin, S. Y. Lee, H. S. Yun, and N. S. Kim, "Speech enhancement based on residual noise shaping," in *Ninth International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, Sep. 2006.

[12] Y. G. Jin, J. W. Shin, and N. S. Kim, "Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. EL228–EL233, Mar. 2017.

[13] H. Kim and J. W. Shin, "Target exaggeration for deep learning-based speech enhancement," *Digital Signal Processing*, vol. 116, pp. 103–109, Sep. 2021.

[14] J. Byun and J. W. Shin, "Monaural speech separation using speaker embedding from preliminary separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2753–2763, Aug. 2021.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, Aug. 2014.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2015, pp. 708–712.

[17] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, Dec. 2015.

[18] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, Aug. 2019.

[19] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K. K. Paliwal, and F. Shang, "Deep residual-dense lattice network for speech enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8552–8559.

[20] A. Nicolson and K. K. Paliwal, "On training targets for deep learning approaches to clean speech magnitude spectrum estimation," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3273–3293, May 2021.

[21] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, Apr. 2020.

[22] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[24] C. V. Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Speech Synthesis Workshop*, Sunnyvale, USA, Sep. 2016, pp. 159–165.

[25] A. Nicolson, "Deep xi dataset," 2020. [Online]. Available: https://dx.doi.org/10.21227/3adt-pb04

[26] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2021, pp. 5749–5753.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] *Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codec*, ITU Recommendation P.862.2, 2007.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, Dallas, Texas, USA, Mar. 2010, pp. 4214–4217.

[30] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.